NEW APPROACHES IN TESTING COMMON ASSUMPTIONS FOR REGRESSIONS WITH MISSING DATA

A Dissertation

by

JUSTIN ANDREW CHOWN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Ursula U. Müller-Harknett
Committee Members,	Raymond J. Carroll
	Jeffrey D. Hart
	Joel Zinn
Head of Department,	Valen E. Johnson

August 2014

Major Subject: Statistics

ABSTRACT

We consider both nonparametric regression and heteroskedastic nonparametric regression models with multivariate covariates and with responses missing at random. The regression function is estimated using a local polynomial smoother, and, when necessary, the scale function is estimated using a combination of local polynomial smoothers. It is shown, for both regression models, that suitable residual-based empirical distribution functions using only the complete cases, i.e. residuals that can actually be constructed from the data, are efficient in the sense of Hájek and Le Cam. In our proofs we derive, more generally, the efficient influence function for estimating an arbitrary linear functional of the error distribution; this covers the distribution function as a special case. Our estimators are shown to admit functional central limit theorems. We do this by applying the transfer principle for complete case statistics, which makes it possible to adapt known results for fully observed data to the case of missing data. Then, we use these residual-based empirical distribution functions to test for normal errors using a martingale transform approach. Small simulation studies are conducted to investigate the performance of these tests. Our results, for the homoskedastic model, show the proposed approach to be comparable to one based on imputation, and, for the heteroskedastic model, the results are sensitive to the estimate of the scale function. Finally, we construct a test for heteroskedasticity using residuals from a nonparametric regression. The approach uses a weighted empirical process and only the completely observed data, and is shown to perform well in certain scenarios. All of the tests considered here are asymptotically distribution free, which means inference based on them does not depend on unknown parameters.

ACKNOWLEDGEMENTS

I would like to thank my advisor Uschi for her careful direction and patient guidance during my studies at Texas A&M University, and her helpful comments during the production of this document. Also, I would like to thank my family and friends for their unending support and encouragement. Lastly, I would like to thank the Taylor & Francis Group for publishing a section of this document in the Journal of Nonparametric Statistics and allowing me to reproduce it in this dissertation.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION*

In this work, we study the nonparametric regression model

$$Y = r(X) + \varepsilon,$$

and the heteroskedastic nonparametric regression model

$$Y = r(X) + \sigma(X)\varepsilon.$$

For both models, the error ε is assumed to be independent of the covariate vector X. Here the function r is called the regression function and the function σ is called the scale function. Nonparametric models are particularly useful for residual-based inference because residuals constructed from these models are usually consistent. This idea is well explored in the literature. For example, Hart (1997) and Ruppert et al. (2009) each review nonparametric and semiparametric approaches. In particular, these authors explore the literature surrounding nonparametric and semiparametric models for estimating the regression and scale functions. Since our work is not directly concerned with either the regression or scale functions, we use local polynomial smoothers to estimate these unknown quantities.

We are interested in the case where the responses Y are missing. In practical applications datasets that contain missing responses are common. Missing information can lead to bias when drawing conclusions, if the missing data is not appropriately handled. Thus, it is important to choose statistical methods that ensure conclusions

*Part of this section is reprinted with permission from "Efficiently estimating the error distribution in nonparametric regression with responses missing at random" by J. Chown and U.U. Müller. *Journal of Nonparametric Statistics*, Volume 25, Issue 3, pages 665-677. Copyright 2013. Citations to Chown and Müller (2013) are given as citations to Section 2.

are not biased.

We make the assumption that responses are *missing at random*. This is a common assumption and is reasonable in many situations (see Chapter 1 of Little and Rubin (2002)). As an example, consider missing responses to a survey question about income. If additional data about medical conditions (X) were available, then we might see that the response probabilities (π) are smaller for subjects diagnosed with depression. In this case the missingness mechanism is ignorable because π depends only on fully observed data X; i.e. it can be estimated from the given data. Further examples of missing data may be found in Tsiatis (2006), Liang et al. (2007), Molenberghs and Kenward (2007), and Efromovich (2011a,b).

In addition, there are many datasets in practice that exhibit heteroskedasticity. For example, consider a study examining the grade point averages of high school students. If additional information on household income (X) were available, then we might see that variation in grade point averages (σ^2) is smaller for households of larger income. In this case, both the mean and variation of grade point averages would depend on household income. This presents a unique challenge to infer conclusions about grade point averages based on these data. Further examples of heteroskedasticity may be found in Asteriou and Hall (2011), Sheather (2009), Vinod (2008) and Greene (2000).

An important tool for making decisions about goodness-of-fit and lack-of-fit is the residual-based empirical distribution function. It is well studied in the literature. For example, Stute (1997) and Khmaladze and Koul (2004, 2009) test parametric hypotheses about the regression function in nonparametric models. Neumeyer and Van Keilegom (2010) study additivity tests in heteroskedastic nonparametric regression. Müller et al. (2012) test for normal errors. The approaches use the residuals from nonparametric regressions, and to study the properties of these tests the error

distribution function is estimated.

In Section 2, we work with the residual-based empirical distribution function of a nonparametric regression to estimate the unknown error distribution function. Our technique extends the approach of Müller et al. (2009) to the missing data case, and we establish this approach as an efficient estimation technique. This estimator uses only the completely observed data; i.e. it is constructed using pairs (X, Y) ignoring the pairs $(X, ?)$. We then apply the test for normal errors of Koul et al. (2012). Here, if an estimator achieves least asymptotic dispersion among those that are consistent with nontrivial limiting behavior, it is called efficient. In particular, these estimators satisfy the Hájek and Le Cam convolution theorem for the special case of a limiting normal distribution (see Schick (1993) for a statement of the theorem).

Interestingly, the efficiency property of our proposed estimator yields that competing estimators will not be able to outperform it in large samples. Specifically, this applies to estimators built using imputation practices. Imputation is a process wherein portions of the incomplete sample are replaced by a suitable estimate $(X, \hat{Y})$, and is usually done in one of two ways. Partial imputation replaces only the incomplete observations, those of the form $(X, ?)$, and preserves the complete observations, those of the form (X, Y). Full imputation replaces the entire sample; i.e. both the incomplete and the complete observations. It is commonly thought that imputation will help to alleviate some of the biases that present in missing data. As a consequence, our results are counter-intuitive to common wisdom.

We work with, in Section 3, another residual-based empirical distribution function of a heteroskedastic nonparametric regression to estimate the unknown error distribution function. It is constructed using only the completely observed data. Our technique extends the approach of Neumeyer and Van Keilegom (2010) to the missing data case, and we establish it as an efficient estimation technique. We find

our approaches satisfies the same conditions of those required of a test for normal errors given in Koul et al. (2012). As a consequence, the results of this section very closely mirror those of the previous section.

In Section 4, we derive a test for heteroskedastic errors using the residuals of a nonparametric regression. Specifically, we study the difference between the nonparametric regression and the heteroskedastic nonparametric regression models. Again, we work with only the completely observed data. The test proposed in this section is inspired by that of Koul et al. (2012) who develop a test for linearity of a semiparametric missing data model. Both approaches are in the spirit of Stute (1997) who considers an empirical process related to the integrated regression function. Each test statistic is constructed by suitably weighting an empirical distribution function. We study a weighted empirical process to obtain the nontrivial limiting behavior of the test statistic.

The manuscript concludes with Section 5. Final remarks are made on the ideas of each problem studied, and reflections are given. These thoughts bring to light some of the deeper questions related to these topics. Many of these questions are considered in practical applications of Statistics.

2. EFFICIENTLY ESTIMATING THE ERROR DISTRIBUTION IN NONPARAMETRIC REGRESSION WITH RESPONSES MISSING AT RANDOM*

In this section we study the nonparametric regression model

$$Y = r(X) + \varepsilon,$$

with the error ε independent of the covariate vector X. We are interested in the case where the responses, Y, are missing; i.e. we observe the sample $(X_1, \delta_1 Y_1, \delta_1), \ldots, (X_n, \delta_n Y_n, \delta_n)$, where δ is an indicator variable which equals one, if Y is observed, and zero, otherwise. Here, we make the assumption that responses are *missing at random* (MAR). This means the probability Y is observed depends only on the covariates,

$$P(\delta = 1|X, Y) = P(\delta = 1|X) = \pi(X).$$

We will refer to the model with responses missing at random as the *MAR model*.

We show the residual-based empirical distribution function $\hat{\mathbb{F}}_c$, defined in equation (2.2) below, to be an efficient estimator of the unknown error distribution function F. This estimator uses only the complete observations, those of the form (X, Y) (the *complete cases*); i.e. the available residuals $\hat{\varepsilon}_{j,c} = Y_j - \hat{r}_c(X_j)$, where $\hat{r}_c$ is a suitable complete case estimator of the regression function. Demonstrating this requires two

*Part of this section is reprinted with permission from "Efficiently estimating the error distribution in nonparametric regression with responses missing at random" by J. Chown and U.U. Müller. *Journal of Nonparametric Statistics*, Volume 25, Issue 3, pages 665-677. Copyright 2013. Citations to Chown and Müller (2013) are given as citations to Section 2.

arguments. First we show that $\hat{\mathbb{F}}_c$ satisfies the uniform stochastic expansion

$$\sup_{t\in\mathbb{R}}\Big|\hat{\mathbb{F}}_c(t) - \frac{1}{N}\sum_{j=1}^{n}\delta_j\mathbf{1}(\varepsilon_j \le t) - f(t)\frac{1}{N}\sum_{j=1}^{n}\delta_j\varepsilon_j\Big| = o_p(n^{-1/2}). \tag{2.1}$$

Here f is the error density and $N = \sum_{j=1}^{n}\delta_j$ is the number of complete cases. Then we show that an estimator of F that admits this expansion is asymptotically *efficient* in the sense of Hájek and Le Cam. To do this we derive, more generally, the efficient influence function for estimating an arbitrary linear functional $E\{h(\varepsilon)\}$. This covers $F(t) = E\{1(\varepsilon \le t)\}$ as a special case. Then we specify the efficient influence function for an efficient estimator of F. We conclude that the estimator $\hat{\mathbb{F}}_c$ with expansion (2.1) is indeed efficient for F.

The first part may be handled easily by employing the *transfer principle* for complete case statistics given in Koul et al. (2012). This principle makes it possible to adapt results for the model where all data are fully observed, the *full model*, to missing data models. In particular, we can use the complete case version $\hat{r}_c$ of the estimator $\hat{r}$ proposed by Müller et al. (2009). They obtain expansion (2.1) for the full model (with all indicators equal to one) using a local polynomial smoother to estimate r.

In order to summarize the main result by Müller et al. (2009) (Theorem 2.1 below) we introduce some notation. Let $i = (i_1, \ldots, i_m)$ be a multi-index, and write $I(k)$ for the set of multi-indices that satisfy $i_1 + \cdots + i_m \le k$. Müller et al. (2009) estimate r by a local polynomial smoother $\hat{r}$ of degree d. It is defined as the component $\hat{\beta}_0$ corresponding to the multi-index $0 = (0, \ldots, 0)$ of a minimizer

$$\hat{\beta} = \operatorname*{arg\,min}_{\beta=(\beta_i)_{i\in I(d)}} \sum_{j=1}^{n}\Big\{Y_j - \sum_{i\in I(d)}\beta_i\psi_i\Big(\frac{X_j - x}{c_n}\Big)\Big\}^2 w\Big(\frac{X_j - x}{c_n}\Big),$$

where

$$\psi_i(x) = \frac{x_1^{i_1}}{i_1!} \cdots \frac{x_m^{i_m}}{i_m!}, \quad x = (x_1, \ldots, x_m) \in \mathbb{R}^m,$$

$w(x) = w_1(x_1) \cdots w_m(x_m)$ is a product of densities, and c_n is a bandwidth.

The estimator $\hat{r}$ permits the desired expansion if the assumptions of Theorem 2.1 (below) are satisfied. This requires, in particular, the regression function r to belong to the Hölder space $H(d, \gamma)$; i.e. it has continuous partial derivatives of order d (or higher), and the partial derivatives of order d are Hölder with exponent γ. The choice of the degree d of the local polynomial smoother will also depend both on smoothness and moment conditions on the error density and on the dimension of the covariate vector. In our simulation study we consider an infinitely differentiable regression function r and a one-dimensional covariate X. This allows us to use a locally linear smoother. Theorem 1 from Müller et al. (2009) is proven under the following assumption on the covariate distribution.

Assumption 2.1. *The covariate vector X is quasi-uniform on the cube $[0,1]^m$; i.e. X has a density which is bounded and bounded away from zero on $[0,1]^m$.*

Theorem 2.1 (Müller et al. (2009), Theorem 1). *Let Assumption 2.1 be satisfied. Suppose the regression function r belongs to $H(d, \gamma)$ with $s = d + \gamma > 3m/2$. Further suppose the error variable to have mean zero, a finite moment of order $\zeta > 4s/(2s-m)$ and a density f that is Hölder with exponent $\xi > m/(2s-m)$. Consider the estimator $\hat{r}$ from above with densities $w_1, \ldots, w_m$ that are $(m+2)$-times continuously differentiable and have compact support $[-1,1]$. Let the bandwidth satisfy $c_n \sim (n \log n)^{-1/(2s)}$. Then, with $\hat{\varepsilon}_j = Y_j - \hat{r}(X_j)$,*

$$\sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{j=1}^{n} \left\{ \mathbf{1}(\hat{\varepsilon}_j \le t) - \mathbf{1}(\varepsilon_j \le t) - \varepsilon_j f(t) \right\} \right| = o_p(n^{-1/2}).$$

We now apply the transfer principle for asymptotically linear statistics given by

Koul et al. (2012) to adapt the results from Theorem 2.1 for the MAR model. The complete case estimator for $F(t)$ is given by

$$\hat{\mathbb{F}}_c(t) = \frac{1}{N}\sum_{j=1}^{n} \delta_j \mathbf{1}(\hat{\varepsilon}_{j,c} \leq t) = \frac{1}{N}\sum_{j=1}^{n} \delta_j \mathbf{1}\{Y_j - \hat{r}_c(X_j) \leq t\}, \tag{2.2}$$

where $\hat{r}_c$ is the complete case version of $\hat{r}$; i.e. the component $\hat{\beta}_{c0}$ of a minimizer

$$\hat{\beta}_c = \underset{\beta=(\beta_i)_{i\in I(d)}}{\arg\min} \sum_{j=1}^{n} \delta_j \Big\{ Y_j - \sum_{i\in I(d)} \beta_i \psi_i\Big(\frac{X_j - x}{c_n}\Big)\Big\}^2 w\Big(\frac{X_j - x}{c_n}\Big). \tag{2.3}$$

Using the transfer principle requires the conditional distribution of (X, Y) given $\delta = 1$ to meet the assumptions on the (unconditional) joint distribution of (X, Y) from Theorem 2.1. In our case it is easy to see this affects only the covariate distribution G: the MAR assumption combined with the independence of X and ε yield that ε and (X, δ) are independent. Hence, the parameters f and r stay the same when switching from the unconditional to the conditional distribution. In particular, the complete case statistic $\hat{\mathbb{F}}_c(t)$ is a consistent estimator for $F(t)$ in the MAR model (since F remains unchanged). Thus we may keep all but one of our assumptions: only Assumption 2.1 must be restated.

ASSUMPTION 2.2. *The conditional distribution of the covariate vector X given $\delta = 1$ is quasi-uniform on the cube $[0, 1]^m$, i.e. it has a density which is bounded and bounded away from zero on $[0, 1]^m$.*

The transfer principle implies the complete case version of the estimator from Theorem 2.1 to satisfy the corresponding expansion (2.1). This expansion is equivalent to

$$\sup_{t\in\mathbb{R}} \Big| \frac{1}{n}\sum_{j=1}^{n} \frac{\delta_j}{E\delta} \big\{ \mathbf{1}(\hat{\varepsilon}_{j,c} \leq t) - \mathbf{1}(\varepsilon_j \leq t) - \varepsilon_j f(t) \big\} \Big| = o_p(n^{-1/2}).$$

Hence we have, uniformly in $t \in \mathbb{R}$,

$$\hat{\mathbb{F}}_c(t) = \frac{1}{n}\sum_{j=1}^{n} \frac{\delta_j}{E\delta}\mathbf{1}(\hat{\varepsilon}_{j,c} \le t) + o_p(n^{-1/2}) = F(t) + \frac{1}{n}\sum_{j=1}^{n} b(\delta_j, \varepsilon_j, t) + o_p(n^{-1/2}),$$

with influence function $b(\delta, \varepsilon, t) = \delta/E\delta\,\{\mathbf{1}(\varepsilon \le t) - F(t) + f(t)\varepsilon\}$. This is indeed the *efficient influence function* for estimating $F(t)$: see Corollary 2.1 below. Now we may state the main result of this section.

THEOREM 2.2. *Consider the nonparametric regression model with responses missing at random. Suppose the assumptions of Theorem 2.1 are satisfied, now with Assumption 2.2 in place of Assumption 2.1. Then the complete case estimator $\hat{\mathbb{F}}_c$ of the error distribution satisfies the stochastic expansion (2.1); i.e.*

$$\sup_{t\in\mathbb{R}} \Big| \frac{1}{N}\sum_{j=1}^{n} \delta_j \big\{ \mathbf{1}(\hat{\varepsilon}_{j,c} \le t) - \mathbf{1}(\varepsilon_j \le t) - \varepsilon_j f(t) \big\} \Big| = o_p(n^{-1/2}).$$

If the error density furthermore fulfills Assumption 2.3, stated in Section 2.1, then $\hat{\mathbb{F}}_c(t)$ is asymptotically efficient, in the sense of Hájek and Le Cam, for estimating $F(t)$, $t \in \mathbb{R}$, with influence function

$$b(\delta, \varepsilon, t) = \frac{\delta}{E\delta}\big\{\mathbf{1}(\varepsilon \le t) - F(t) + f(t)\varepsilon\big\}.$$

Remark 2.1. If the transfer principle were not available, then the expansion in Theorem 2.2 could be derived by mimicking the (rather elaborate) proofs of Lemma 1 in Müller et al. (2009) and of Theorem 2.2 in Müller et al. (2007): who estimate the error distribution in a general semiparametric regression model. Our arguments are essentially the same – what is new now is the presence of indicators. The approach is as follows. Analogously to (Müller et al., 2009, equation (1.4)), one derives an

approximation $\hat{a}_c(x)$ of the difference $\hat{r}_c(x) - r(x)$,

$$\sup_{x\in\mathbb{R}} |\hat{r}_c(x) - r(x) - \hat{a}_c(x)| = o_p(n^{-1/2}). \tag{2.4}$$

Note, the statements $\hat{\varepsilon}_{j,c} \le t$ and $\varepsilon_j \le t + \hat{r}_c(x) - r(x)$ are equivalent. Now use this and (2.4) and replace the two empirical distribution functions in the formula by their respective expectations (cf. (Müller et al., 2007, proof of Theorem 2.2)); i.e. replace $\hat{\mathbb{F}}_c$ by $F_{\hat{a}_c}(t)$ and $N^{-1}\sum_{j=1}^n \delta_j \mathbf{1}(\varepsilon_j \le t)$ by $F(t)$. This gives

$$\sup_{t\in\mathbb{R}} \Big| \frac{1}{n}\sum_{j=1}^{n} \frac{\delta_j}{E\delta} \{\mathbf{1}(\hat{\varepsilon}_{j,c} \le t) - \mathbf{1}(\varepsilon_j \le t)\} - \{F_{\hat{a}_c}(t) - F(t)\} \Big| = o_p(n^{-1/2}),$$

where

$$F_a(t) = E\Big[\frac{\delta_j}{E\delta}\mathbf{1}\{\varepsilon \le t + a(X)\}\Big] = E[\mathbf{1}\{\varepsilon \le t + a(X)\}|\delta = 1] = \int F\{t + a(x)\}\, G_1(dx)$$

with G_1 denoting the conditional distribution of X given $\delta = 1$. Here $F(t)$ is the expectation of the second term of the sum; i.e. $F(t) = F_a(t)$ for $a = 0$. A Taylor expansion applied to $F_{\hat{a}_c}(t) - F(t)$ in the above expansion yields

$$\sup_{t\in\mathbb{R}} \Big| \frac{1}{n}\sum_{j=1}^{n} \frac{\delta_j}{E\delta} \big\{\mathbf{1}(\hat{\varepsilon}_{j,c} \le t) - \mathbf{1}(\varepsilon_j \le t)\big\} - f(t) \int \hat{a}_c(x) G_1(dx) \Big| = o_p(n^{-1/2}).$$

The desired expansion now follows from this combined with

$$\int \hat{a}_c(x) G_1(dx) = \frac{1}{n}\sum_{j=1}^{n} \frac{\delta_j}{E\delta}\varepsilon_j + o_p(n^{-1/2}).$$

The last approximation is the complete case version of equation (1.3) in Müller et al. (2009). It can be verified by inspecting the proof of Lemma 1 in that paper, where properties of local polynomial smoothers are derived. Keep in mind that our estimators are constructed from the complete cases (equation (2.3) above), which explains the indicators in the above formula.

Note, the uniform expansion implies a *functional central limit theorem.* Also note, the efficiency property of our proposed simple estimator $\hat{\mathbb{F}}_c$ yields that competing imputation type estimators will not be able to outperform it in large samples. Below, we illustrate this result with simulations for two examples. The first example demonstrates the efficiency of the complete case estimator $\hat{\mathbb{F}}_c$ by comparing it with a 'tuned' estimator using an imputation technique similar to one studied by González-Manteiga and Pérez-González (2006). For our second example, we perform simulations similar to those in Müller et al. (2012), who use a martingale transform approach to test for normal errors in the full model. The test statistics involve the estimators from the first example.

2.1 Efficiency

Now we calculate the efficient influence function for estimating the functional $E\{h(\varepsilon)\}$ using observations $(X_i, \delta_i Y_i, \delta_i)$, $i = 1, \ldots, n$. First, we follow the arguments of Müller et al. (2006), who study efficient estimation of general differentiable functionals with data of the above form. We summarize their main arguments and refer to that paper for more details. Then, we focus on the functional $F(t)$, which Müller et al. (2004) study in the full model. This allows us to adapt parts of their proofs to the MAR model considered here.

No assumption of a *parametric* model (finite dimensional) for the regression function or for the distribution of the observations is made. Thus, the parameter set Θ of the statistical model includes a family of covariate distributions, $\mathscr{G}$; a family of error distributions, $\mathscr{F}$; a set of regression functions, $\mathscr{R}$; and a family of response probability distributions, $\mathscr{B}$; i.e. $\Theta = \mathscr{G} \times \mathscr{F} \times \mathscr{R} \times \mathscr{B}$. We impose the following assumptions:

ASSUMPTION 2.3. *The error density f is absolutely continuous with almost everywhere derivative f' and finite Fisher information $J = \int \ell^2(z) f(z) dz$, where $\ell = -f'/f$ denotes the score function.*

Since the construction of the efficient influence function utilizes the directional information in Θ, we will now identify the set $\dot{\Theta}$ of all perturbations related to the statistical model, which may be thought of as directions. The joint distribution $P(dx, dy, dz)$ depends on the marginal distribution $G(dx)$ of X, the conditional probability $\pi(x)$ that δ equals one given $X = x$, and the conditional distribution $Q(x, dy)$ of Y given $X = x$. Formally, we have

$$P(dx, dy, dz) = G(dx) B_{\pi(x)}(dz) \left\{ z Q(x, dy) + (1 - z) \delta_0(dy) \right\},$$

where $B_p = p\delta_1 + (1-p)\delta_0$ denotes the Bernoulli distribution with parameter p and δ_t is the Dirac measure at t. Now consider perturbations G_{nu}, π_{nw} and Q_{nv} of G, π and Q, respectively, that are *Hellinger differentiable* in the following sense:

$$\int \left\{ n^{1/2} \big(dG_{nu}^{1/2} - dG^{1/2} \big) - \frac{1}{2} u dG^{1/2} \right\}^2 \to 0,$$

$$\int \int \left[n^{1/2} \big\{ dB^{1/2}_{\pi_{nw}(x)} - dB^{1/2}_{\pi(x)} \big\} - \frac{1}{2} \{ \cdot - \pi(x) \} w(x) dB^{1/2}_{\pi(x)} \right]^2 G(dx) \to 0,$$

$$\int \int \left[(n^{1/2} \big\{ dQ^{1/2}_{nv}(x, \cdot) - dQ^{1/2}(x, \cdot) \big\} - \frac{1}{2} v(x, \cdot) dQ^{1/2}(x, \cdot) \right]^2 G_1(dx) \to 0,$$

with G_1 as the conditional distribution of X given that $\delta = 1$. This requires that u belongs to $\mathcal{L}_{2,0}(G)$; i.e. $u \in \mathcal{L}_2(G)$ and $\int u \, dG = 0$. Further, we require that w belongs to

$$\mathcal{L}_2(G_\pi) = \left\{ w \in \mathcal{L}_2(G) : \int w^2(x) \pi(x) \{ 1 - \pi(x) \} dG(x) < \infty \right\}$$

with $G_\pi(dx) = \pi(x)\{1-\pi(x)\}G(dx)$, and that v belongs to

$$\mathcal{V}_0 = \left\{ v \in \mathcal{L}_2(Q \otimes G_1) : \int v(x,y) dQ(x,dy) = 0 \right\}.$$

Note that models for G_1, π and Q will imply further restrictions on the perturbations in order to satisfy those model assumptions. So u, w and v must be restricted to subspaces of $\mathcal{L}_{2,0}(G)$, $\mathcal{L}_2(G_\pi)$ and $\mathcal{V}_0$, respectively. In this work no model assumptions on G and π have been made. Hence, we only need to identify the appropriate subspace $\mathcal{V}$ of $\mathcal{V}_0$. Since the covariates and the errors are assumed to be independent, we may write $Q(x,dy) = f\{y-r(x)\}dy$. With this notation the constraint on $v \in \mathcal{V}_0$ states that $\int v(x,y) f\{y-r(x)\}dy = 0$. In order to derive the explicit form of $\mathcal{V}$ we introduce perturbations s and t of the unknown functions f and r and write

$$Q_{nv}(x,dy) = Q_{nst}(x,dy) = f_{ns}(y-r_{nt})dy,$$

where $f_{ns}(z) = f(z)\{1+n^{-1/2}s(z)\}$ and $r_{nt}(x) = r(x) + n^{-1/2}t(x)$ for $s \in \mathcal{S}$ and $t \in \mathcal{T}$. Here

$$\mathcal{S} = \left\{ s \in \mathcal{L}_2(F) : \int s(z)f(z)dz = 0, \int zs(z)f(z)dz = 0 \right\},$$

which comes from the requirement for the perturbed density f_{ns} to integrate to one and have mean zero. We may take $\mathcal{T} = \mathcal{L}_2(G_1)$ since we do not assume a parametric form for r. In the following we will write "$\doteq$" to denote asymptotic equivalence; i.e. equality up to an additive term of order $o_p(n^{-1/2})$. As in Müller (2009), who considers a parametric regression function (nonlinear), we have

$$\begin{aligned} f_{ns}(y-r_{nt}(x)) &= f\{y-r_{nt}(x)\}\left[1+n^{-1/2}s\{y-r_{nt}(x)\}\right] \\ &= f\{y-r(x)-n^{-1/2}t(x)\}\left[1+n^{-1/2}s\{y-r(x)-n^{-1/2}t(x)\}\right] \\ &\doteq f\{y-r(x)\}\Big(1+n^{-1/2}\big[s\{y-r(x)\}+\ell\{y-r(x)\}t(x)\big]\Big). \end{aligned}$$

Hence $Q_{nst}(x, dy) \doteq f\{y - r(x)\}\Big(1 + n^{-1/2}\big[s\{y - r(x)\} + \ell\{y - r(x)\}t(x)\big]\Big)$ and $\mathcal{V}$ has the form

$$\mathcal{V} = \Big\{v(x, y) = s\{y - r(x)\} + \ell\{y - r(x)\}t(x) : s \in \mathcal{S}, t \in \mathcal{T}\Big\}.$$

Thus we construct $\dot{\Theta}$ as the set containing all possible Hellinger perturbations of the statistical model parameters, or just $\dot{\Theta} = \mathcal{L}_{2,0}(G) \times \mathcal{S} \times \mathcal{L}_2(G_1) \times \mathcal{L}_2(G_\pi)$. The perturbed distribution $P_{n\gamma}$, with $\gamma = (u, s, t, w)$ in $\dot{\Theta}$, of the observation $(X, \delta Y, \delta)$ is then

$$P_{n\gamma}(dx, dy, dz) \doteq G_{nu}(dx)B_{\pi_{nw(x)}}(dz)\left\{zQ_{nst}(x, dy) + (1 - z)\delta_0(dy)\right\}.$$

It follows that $P_{n\gamma}$ is Hellinger differentiable with tangent

$$d_\gamma(X, \delta Y, \delta) = u(X) + \delta\{s(\varepsilon) + \ell(\varepsilon)t(X)\} + \{\delta - \pi(X)\}w(X).$$

The efficient influence function of a differentiable functional is characterized by its canonical gradient, which is defined as an orthogonal projection of a gradient onto the tangent space. We take the tangent space T as the closure of the linear subspace formed by d_γ. Since d_γ is a sum of orthogonal elements we may write

$$T = \mathcal{L}_{2,0}(G) \oplus \{(\delta - \pi(X))w(X) : w \in \mathcal{L}_2(G_\pi)\} \oplus \{\delta v(X, Y) : v \in \mathcal{V}\}.$$

We are interested in the linear functional $E\{h(\varepsilon)\}$. In order to specify a gradient of $E\{h(\varepsilon)\}$ we need the directional derivative $\gamma_h \in \dot{\Theta}$ of $E\{h(\varepsilon)\}$, which is characterized by a limit as follows. As in Müller et al. (2004), we have, for every $s \in \mathcal{S}$,

$$\lim_{n\to\infty} n^{1/2}\Big[\int h(z)f_{ns}(z)\,dz - E\{h(\varepsilon)\}\Big] = E\{h(\varepsilon)s(\varepsilon)\} = E\{h_0(\varepsilon)s(\varepsilon)\},$$

with h_0 given as the projection of h onto $\mathcal{S}$:

$$h_0(z) \;=\; h(z) - \int h\,dF - \frac{z}{\sigma^2}\int x h(x)\,dF(x),$$

where σ^2 denotes the error variance. Hence $E\{h(\varepsilon)\}$ is differentiable with directional derivative $\gamma_h = (0, h_0, 0, 0)$ and gradient $h_0(\varepsilon)$. By the convolution theorem (see, for example, (Schick, 1993, Section 2)), the unique canonical gradient $g^*(X, \delta Y, \delta)$ is obtained as the orthogonal projection of $h_0(\varepsilon)$ onto the tangent space T. Hence it must be of the form

$$g^*(X, \delta Y, \delta) = u^*(X) + \delta\{s^*(\varepsilon) + \ell(\varepsilon)t^*(X)\} + \{\delta - \pi(X)\}w^*(X) \tag{2.5}$$

and is characterized by

$$E\{h_0(\varepsilon)s(\varepsilon)\} = E\{g^*(X, \delta Y, \delta)d_\gamma(X, \delta Y, \delta)\} \tag{2.6}$$

for every $\gamma \in \dot{\Theta}$. A straightforward calculation yields for the right-hand side of (2.6):

$$\begin{aligned}
&E\{g^*(X, \delta Y, \delta)d_\gamma(X, \delta Y, \delta)\} \\
&= E\big\{u^*(X)u(X)\big\} + E\delta E\big\{s^*(\varepsilon)s(\varepsilon)\big\} + E\big\{\ell_0(\varepsilon)s^*(\varepsilon)\big\}E\big\{\pi(X)t(X)\big\} \\
&\quad + E\big\{\ell_0(\varepsilon)s(\varepsilon)\big\}E\big\{\pi(X)t^*(X)\big\} + JE\big\{t^*(X)t(X)\big\} \\
&\quad + E\big[\pi(X)\{1 - \pi(X)\}w^*(X)w(X)\big],
\end{aligned}$$

where $\ell_0(\varepsilon)$ is the projection of $\ell(\varepsilon)$ onto $\mathcal{V}$, that is $\ell_0(\varepsilon) = \ell(\varepsilon) - \varepsilon/\sigma^2$. For convenience, we introduce the quantity J_0 which is calculated analogously to J as

$$J_0 = \int \ell_0^2\,dF = \int \Big\{\ell(z) - \frac{z}{\sigma^2}\Big\}^2 dF(z) = J - \frac{1}{\sigma^2}.$$

From (2.6) it is easy to see that $u^* = w^* = 0$. Setting $u = t = w = 0$ in (2.6) we

obtain

$$\int h_0 s\, dF = E\delta \int s s^*\, dF + \int \ell_0 s\, dF \int \pi t^*\, dG$$

for all s. This gives $s^*(z) = (E\delta)^{-1}\{h_0(z) - \ell_0(z) \int \pi t^*\, dG\}$. Now set $u = s = w = 0$ in (2.6) and insert s^* to get

$$\begin{aligned} 0 &= \int \ell_0 s^*\, dF \int \pi t\, dG + J \int \pi t^* t\, dG = \int \ell_0 s^*\, dF E\delta \int t\, dG_1 + JE\delta \int t^* t\, dG_1 \\ &= \int h_0 \ell_0\, dF \int t\, dG_1 - J_0 E\delta \int t\, dG_1 \int t^*\, dG_1 + JE\delta \int t^* t\, dG_1 \end{aligned}$$

for all $t \in \mathcal{L}_2(G_1)$. Now consider $\mathcal{L}_2(G_1)$ written (as in Müller et al. (2004)) as an orthogonal sum of functions with mean 0 and of constants, i.e.
$\mathcal{L}_2(G_1) = \mathcal{L}_{2,0}(G_1) \oplus [1]$, which means that we can write $t = (t - \int t\, dG_1) + \int t\, dG_1$. The above equation now becomes

$$\begin{aligned} 0 = JE\delta \int \big(t - \int t\, dG_1\big)\big(t^* - \int t^*\, dG_1\big)\, dG_1 \\ + \int h_0 \ell_0\, dF \int t\, dG_1 + \frac{E\delta}{\sigma^2} \int t\, dG_1 \int t^*\, dG_1 \end{aligned}$$

for all $t \in \mathcal{L}_2(G_1)$. This yields

$$t^* - \int t^*\, dG_1 = 0, \quad \int t^*\, dG_1 = -\sigma^2 (E\delta)^{-1} \int h_0 \ell_0\, dF$$

and, thus, $t^* = -\sigma^2 (E\delta)^{-1} \int h_0 \ell_0\, dF$. Combining the above calculations we obtain the following result:

LEMMA 2.1. *The canonical gradient of $E\{h(\varepsilon)\}$ is $g^*(X, \delta Y, \delta)$ and characterized by $(0, s^*, t^*, 0)$, where*

$$s^*(z) = \frac{1}{E\delta}\big[h_0(z) + \sigma^2 E\{h_0(\varepsilon)\ell_0(\varepsilon)\}\ell_0(z)\big] \quad \text{and} \quad t^* = -\frac{\sigma^2}{E\delta} E\{h_0(\varepsilon)\ell_0(\varepsilon)\},$$

with $\sigma^2 = E(\varepsilon^2)$, $h_0(\varepsilon) = h(\varepsilon) - \int h\, dF - \varepsilon\sigma^{-2} \int z h(z)\, dF(z)$ and $\ell_0(\varepsilon) = \ell(\varepsilon) - \varepsilon/\sigma^2$.

An estimator $\hat{\mu}$ of $E\{h(\varepsilon)\}$ is efficient, in the sense of Hájek and Le Cam, if it is asymptotically linear with influence function equal to the canonical gradient $g^*(X, \delta Y, \delta)$ that characterizes $E\{h(\varepsilon)\}$; i.e. if

$$n^{1/2}\big\{\hat{\mu} - E\{h(\varepsilon)\}\big\} = n^{-1/2}\sum_{i=1}^{n} g^*(X_i, \delta_i Y_i, \delta_i) + o_p(1).$$

A straightforward calculation using this combined with Lemma 2.1 and formula (2.5) yields:

THEOREM 2.3. *Consider the nonparametric regression model with responses missing at random. An efficient estimator $\hat{\mu}$ of $E\{h(\varepsilon)\}$ must satisfy the expansion*

$$n^{1/2}\big[\hat{\mu} - E\{h(\varepsilon)\}\big] = n^{-1/2}\sum_{i=1}^{n} \frac{\delta_i}{E\delta}\big[h(\varepsilon_i) - E\{h(\varepsilon)\} - \varepsilon_i E\{\ell(\varepsilon)h(\varepsilon)\}\big] + o_p(1).$$

Remark 2.2. Müller et al. (2004) construct residual-based estimators $n^{-1}\sum_{i=1}^{n} h(\hat{\varepsilon}_i)$ for estimating $E\{h(\varepsilon)\}$ in the full model. In their Section 2 they give conditions for the i.i.d. representation

$$n^{-1/2}\sum_{i=1}^{n} h(\hat{\varepsilon}_i) = n^{-1/2}\sum_{i=1}^{n}[h(\varepsilon_i) - E\{h'(\varepsilon)\}\varepsilon_i] + o_p(1),$$

which characterizes an efficient estimator. (For simplicity, we assume in this remark that h is differentiable.) Note that $E\{h'(\varepsilon)\} = E\{\ell(\varepsilon)h(\varepsilon)\}$. Using the transfer principle, we find the complete case versions of their estimators to satisfy the expansion from the previous theorem. Therefore, they are efficient in the MAR model.

The function $h(\varepsilon) = \mathbf{1}(\varepsilon \leq t)$ is of particular interest since many statistical methods are residual-based and require estimation of the error distribution function. Using Theorem 2.3 with this particular $h(\varepsilon)$, we obtain an expansion for the residual-based empirical distribution function, given in the following corollary.

COROLLARY 2.1. *Consider the nonparametric regression model with responses miss-*

ing at random. An estimator $\hat{F}$ of the error distribution function F is efficient if it satisfies the expansion

$$n^{1/2}\big\{\hat{F}(t) - F(t)\big\} = n^{-1/2}\sum_{i=1}^{n}\frac{\delta_i}{E\delta}\big\{\mathbf{1}(\varepsilon_i \le t) - F(t) + \varepsilon_i f(t)\big\} + o_p(1).$$

This is the expansion of the complete case estimator $\hat{\mathbb{F}}_c$ from Section 2, which completes the proof of Theorem 2.2.

2.2 Simulation studies

To conclude this section we present a brief simulation study of the previous results. We also apply a goodness-of-fit test for normal errors to the residuals. For both examples we assume a nonparametric regression model as before, $Y = r(X)+\varepsilon$. In order to depict the nonparametric nature of the regression function r, we choose for the simulations

$$r(x) = x^3 - x^2 + x + \cos\Big(\frac{3\pi}{2}x\Big).$$

The covariates were generated from a uniform distribution and the errors from a normal distribution: $X_i \sim U(-1,1)$ and $\varepsilon_i \sim N(0,1)$ for $i = 1,\ldots,n$; see Figure 2.1 which shows a scatterplot of a simulated dataset. Finally, the indicators δ_i have a Bernoulli($\pi(x)$) distribution, with $\pi(x) = P(\delta = 1|X = x)$. For the simulations we use the logistic distribution function for $\pi(x)$, with a mean of zero and scale parameter of one,

$$\pi(x) \;=\; \frac{1}{1+e^{-x}}.$$

Therefore, the mean amount of missing data is around 50% and ranges between 27% and 73%. For the above choices the assumptions of Theorem 2.2 are satisfied. We work with $d = 1$, the local linear smoother, with bandwidth $c_n = 1.25\{n\log(n)\}^{-1/4}$.

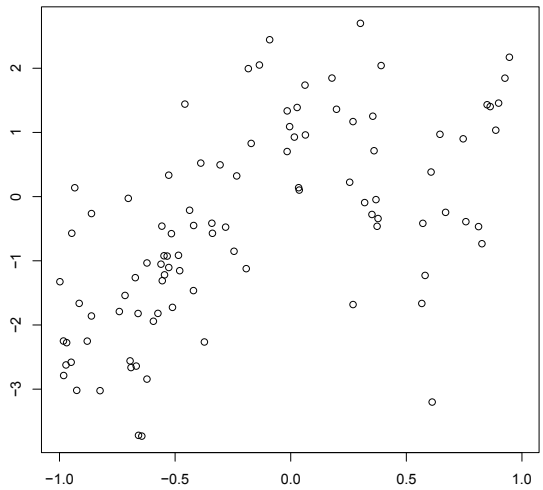

Figure 2.1: $r(x) = x^3 - x^2 + x + \cos\left(\frac{3\pi}{2}x\right)$, $-1 \leq x \leq 1$, with $N(0,1)$ errors

2.2.1 Example 1: Simulation of asymptotic mean squared error

We consider two estimators of the error distribution function. The first estimator is the proposed complete case estimator $\hat{\mathbb{F}}_c$ and the second is a 'tuned' version of $\hat{\mathbb{F}}_c$ that utilizes an imputation technique. Similar to González-Manteiga and Pérez-González (2006), we take the initial local polynomial complete case estimator $\hat{r}_c$ (see equation (2.3)) to produce the completed sample $(X_i, \hat{Y}_i)$ for $i = 1, \ldots, n$. We chose $\hat{Y}_i = \hat{r}_c(X_i)$ for each $i = 1, \ldots, n$. This is a variation of the approach of González-Manteiga and Pérez-González who work with $\hat{Y}_i = \delta_i Y_i + (1-\delta_i)\hat{r}_c(X_i)$; i.e. a "partial imputation" technique. A new local polynomial fit, $\hat{r}^*(\cdot)$, is then constructed from the completed sample. If Y is observed, then we can compute adjusted residuals of

the form $\hat{\varepsilon}^* = Y - \hat{r}^*(X)$. Using these residuals we obtain the new tuned estimator

$$\hat{\mathbb{F}}_\iota(t) = N^{-1} \sum_{j=1}^{n} \delta_j \mathbf{1}\big(\hat{\varepsilon}_j^* \leq t\big).$$

From the previous sections we know the complete case estimator $\hat{\mathbb{F}}_c$ to be an efficient estimator of the error distribution function. The discussion in Remark 2.1 suggests the tuned estimator $\hat{\mathbb{F}}_\iota$ to also be efficient; i.e. both estimators are asymptotically equivalent: we expect that $\hat{\mathbb{F}}_\iota$ can be expanded in the same way as $\hat{\mathbb{F}}_c$,

$$\sup_{t\in\mathbb{R}} \Big| \frac{1}{n} \sum_{j=1}^{n} \frac{\delta_j}{E\delta} \big\{ \mathbf{1}(\hat{\varepsilon}_j^* \leq t) - \mathbf{1}(\varepsilon_j \leq t) \big\} - f(t) \int \hat{a}^*(x) G_1(dx) \Big| = o_p(n^{-1/2}),$$

where $\hat{a}^*(x)$ is now an approximation of the difference $\hat{r}^*(x) - r(x)$ (cf. equation (2.4) in Remark 2.1). The term involving the integral can be written as

$$f(t) \int \hat{a}^*(x) G_1(dx) = f(t) \int \hat{a}_c(x) G_1(dx) + f(t) \int \{\hat{a}^*(x) - \hat{a}_c(x)\} G_1(dx),$$

with the last term being asymptotically negligible since $\hat{a}^*(x) - \hat{a}_c(x)$ is the difference $\hat{r}^*(x) - \hat{r}_c(x)$ of two consistent estimators of $r(x)$. The arguments from Remark 2.1 would then yield

$$\sup_{t\in\mathbb{R}} \Big| \frac{1}{N} \sum_{j=1}^{n} \delta_j \big\{ \mathbf{1}(\hat{\varepsilon}_j^* \leq t) - \mathbf{1}(\varepsilon_j \leq t) - \varepsilon_j f(t) \big\} \Big| = o_p(n^{-1/2}).$$

Indicating both $\hat{\mathbb{F}}_c$ and $\hat{\mathbb{F}}_\iota$ have the same asymptotic expansion.

In order to further check the conjecture that both estimators are asymptotically equivalent, we conducted a simulation study using 1000 trials. We considered four sample sizes and five different values of t at which the error distribution function was evaluated. The findings are summarized in Table 2.1. Note that we also implemented another estimator, which uses partial imputation to complete the sample as suggested by González-Manteiga and Pérez-González. Since our approach performed slightly

Asymptotic mean squared error (MSE)										
	$t=-1.5$		$t=-1$		$t=0$		$t=1$		$t=1.5$	
n	$\hat{\mathbb{F}}_c$	$\hat{\mathbb{F}}_\iota$	$\hat{\mathbb{F}}_c$	$\hat{\mathbb{F}}_\iota$	$\hat{\mathbb{F}}_c$	$\hat{\mathbb{F}}_\iota$	$\hat{\mathbb{F}}_c$	$\hat{\mathbb{F}}_\iota$	$\hat{\mathbb{F}}_c$	$\hat{\mathbb{F}}_\iota$
50	0.1141	0.0987	0.2705	0.2087	0.1702	0.1884	0.2865	0.2220	0.1179	0.1009
250	0.1018	0.0930	0.1800	0.1634	0.2021	0.2071	0.2022	0.1972	0.1201	0.1165
1000	0.0991	0.0945	0.1668	0.1625	0.1865	0.1997	0.1706	0.1780	0.1000	0.1008
10000	0.0925	0.0920	0.1567	0.1537	0.2068	0.2274	0.1690	0.1752	0.0953	0.0975
true	0.0911	–	0.1498	–	0.1816	–	0.1498	–	0.0911	–

Table 2.1: Simulated and true asymptotic MSE

better, we report only the results for our version of $\hat{\mathbb{F}}_\iota$ which is based on $\hat{Y}_i = \hat{r}_c(X_i)$; i.e. both observed and unobserved responses are imputed. For the second smoothing step we chose the same bandwidth as in the first step, $c_n = 1.25\{n\log(n)\}^{-1/4}$.

These results show the simulated MSE (multiplied by n) of our efficient estimator to be close to the true asymptotic MSE (which equals the asymptotic variance and can be calculated using Corollary 2.1). We also see the asymptotic MSE estimates of $\hat{\mathbb{F}}_\iota$ to behave in a similar way to those of $\hat{\mathbb{F}}_c$, in particular for large sample sizes. This provides further evidence of the two approaches being asymptotically equivalent. The simulated MSE's of $\hat{\mathbb{F}}_\iota$, however, more closely match the true asymptotic MSE across values of t at low sample sizes. This could be a second order effect and we believe the most likely explanation is that $\hat{\mathbb{F}}_\iota$ can be regarded as an enhanced version of $\hat{\mathbb{F}}_c$. However, when $t = 0$ both estimators, $\hat{\mathbb{F}}_\iota$ and $\hat{\mathbb{F}}_c$, perform very similarly for all sample sizes. Since this value of t is also the mode of the distribution, we believe the tuning technique using imputation is least helpful in this case.

2.2.2 Example 2: Simulating a goodness-of-fit test for normal errors

We now consider a test proposed by Müller et al. (2012) for the full model with multivariate covariates. This test was also examined by Koul et al. (2012) in the MAR model with a one-dimensional covariate, but without simulations. Both articles study versions of a martingale transform test developed by Khmaladze and Koul (2009).

Under the null hypothesis, these tests tend in distribution to $\sup_{t\in[0,1]}|B(t)|$, with $B(t)$ the standard Brownian motion; i.e. they are asymptotically distribution free. This is very useful since the corresponding complete case statistics have the same limiting distributions in this case, which is a consequence of the transfer principle. So the decision rule remains unchanged in the MAR model. For example, setting the level of the test to 0.05, we reject H_0 if the test statistic exceeds 2.2414, the upper 5% quantile of the distribution of $\sup_{t\in[0,1]}|B(t)|$.

Writing $\phi(x)$ for the density of the $N(0,1)$ distribution and σ^2 for the error variance, the null hypothesis of normal errors is

$$H_0 : \exists\sigma > 0 \quad f(x) = \frac{1}{\sigma}\phi\Big(\frac{x}{\sigma}\Big), \quad x \in \mathbb{R}.$$

In order to introduce the test statistic, T_c, consider

$$H(t) = \int_{-\infty}^{t} h^T(x)\Gamma^{-1}(x)\phi(x)dx,$$

with $\Gamma(x) = \int_x^\infty h(z)h^T(z)\phi(z)dz$ and $h(x) = (1, -\phi'(x)/\phi(x), -(x\phi(x))'/\phi(x))^T$ (see Müller et al. (2012) and Koul et al. (2012) for an explicit form of $\Gamma(x)$ and for more details). Following Koul et al. (2012) we have the test statistic

$$T_c = \sup_{t\in\mathbb{R}}\Big|\frac{1}{\sqrt{N}}\sum_{j=1}^{n}\delta_j\big\{\mathbf{1}(\hat{Z}_{j,c}\le t) - H(t\wedge\hat{Z}_{j,c})h(\hat{Z}_{j,c})\big\}\Big|.$$

Note that this statistic is based on our proposed estimator $\hat{\mathbb{F}}_c$ but with *scaled* residuals $\hat{Z}_{j,c} = \hat{\varepsilon}_{j,c}/\hat{\sigma}_c$, where $\hat{\sigma}_c$ is the complete case version of the residual-based empirical estimator; i.e. $\hat{\sigma}_c = \sqrt{\hat{\sigma}_c^2}$ with

$$\hat{\sigma}_c^2 = \frac{1}{N}\sum_{j=1}^{n}\delta_j\hat{\varepsilon}_{j,c}^2 = \frac{1}{N}\sum_{j=1}^{n}\delta_j\{Y_j - \hat{r}_c(X_j)\}^2.$$

Under the MAR assumption ε and δ are independent. Hence $\hat{\sigma}_c^2$ is a consistent

Test for normal errors								
	$N(0,2)$		$\chi^2_1 - 1$		t_4		Laplace$(0,2)$	
n	T_c	T_ι	T_c	T_ι	T_c	T_ι	T_c	T_ι
50	0.022	0.025	0.489	0.535	0.099	0.108	0.095	0.119
200	0.030	0.028	1.000	1.000	0.457	0.463	0.459	0.483

Table 2.2: Simulated level (N(0,2) figures) and power for T_c and T_ι

estimator of $\text{Var}(\varepsilon|\delta = 1) = \text{Var}(\varepsilon) = \sigma^2$.

We are interested in studying the performance of T_c in the MAR model, and also wish to compare it with the corresponding statistic T_ι that is based on the tuned estimator $\hat{\mathbb{F}}_\iota$; i.e. T_ι has exactly the same form as T_c but with all $\hat{\varepsilon}_{j,c}$ replaced by the adjusted residuals $\hat{\varepsilon}^*_j = Y_j - \hat{r}^*(X_j)$. For the simulations we consider the same scenario as in the previous example, but now also admit some other models for the error distribution. First we look at the $N(0,2)$ distribution to allow verification of the (5%) level of the test. For the power considerations we generated errors from a mean shifted $\chi^2(1)$ distribution, a $t(4)$ distribution and a Laplace distribution with mean 0 and variance 2. The simulation study is based on 1000 runs and samples of sizes 50 and 200.

Table 2.2 shows that when the errors are normally distributed (and the null hypothesis is true) the test using T_c rejects the null hypothesis 2.2% of the time for samples of size 50, and 3% of the time for samples of size 200. This indicates the test using T_c is slightly conservative. Turning to T_ι we see similar conservative behavior: here the hypothesis of normality is rejected 2.5% and 2.8% of the time for sample sizes 50 and 200, respectively. When the null hypothesis is not true, the power figures are fairly close for both tests. The test using T_ι seems to be more powerful for low sample sizes. We find the differences are less pronounced for the larger sample size of

200 suggesting the two tests are asymptotically equivalent – which is what we would expect given the discussion and the simulation results in the previous example. In conclusion, both test procedures have similar performance. The test based on T_c appears to be the better choice for moderately large (or large) samples as it is easier to implement.

3. TESTING FOR NORMAL ERRORS IN HETEROSKEDASTIC NONPARAMETRIC REGRESSION WITH MISSING DATA

In this section we study the heteroskedastic nonparametric regression model

$$Y = r(X) + \sigma(X)\varepsilon,$$

with the error ε independent of the covariate vector X. This model is closely related to that studied in the previous section and so many results will be familiar. Here we will estimate the two functions r and σ with nonparametric techniques. We are interested in the case where the responses, Y, are missing; i.e. we observe a sample $(X_1, \delta_1 Y_1, \delta_1), \ldots, (X_n, \delta_n Y_n, \delta_n)$, where δ is an indicator variable taking values one, when Y is observed, and zero, otherwise. In this work, we assume the responses are missing at random and again refer to the model with reponses missing at random as the MAR model (see Section 2 for details).

The arguments in this section will show the residual-based empirical distribution function, $\hat{\mathbb{F}}_c$, defined in equation (3.2) below, to be an efficient estimator of the unknown error distribution function F. Similar to the estimator in the previous section, this estimator uses only the completely observed data; i.e. the available residuals $\hat{\varepsilon}_{i,c} = \{Y_i - \hat{r}_c(X_i)\}/\hat{\sigma}_c(X_i)$, where $\hat{r}_c$ is a suitable complete case estimator for the regression function r and $\hat{\sigma}_c$ is a suitable complete case estimator of the scale function σ. Demonstrating this fact will require two arguments. For the first argument, we show that $\hat{\mathbb{F}}_c$ satisfies the following uniform stochastic expansion

$$\sup_{t\in\mathbb{R}} \left| \hat{\mathbb{F}}_c - \frac{1}{N}\sum_{j=1}^{n} \delta_j \mathbf{1}(\varepsilon_j \le t) - f(t)\frac{1}{N}\sum_{j=1}^{n} \delta_j \Big\{ \varepsilon_j + \frac{t}{2}(\varepsilon_j^2 - 1) \Big\} \right| = o_p(n^{-1/2}). \quad (3.1)$$

In the above display, f is the error density and $N = \sum_{j=1}^{n} \delta_j$ is the number of complete cases.

We will employ the transfer principle for complete case statistics given in Koul et al. (2012) as was done in Section 2. To do this, we must first consider the full model. Neumeyer and Van Keilegom (2010) investigate this case using estimators of the regression and scale functions which are based on local polynomial smoothers, and obtain expansion (3.1). This means that we may use complete case versions $\hat{r}_c$ and $\hat{\sigma}_c$ by identifying suitable complete case local polynomial smoothers.

To summarize the result of Neumeyer and Van Keilegom (2010) (written as Theorem 3.1 below) we will introduce some notation. First, Neumeyer and Van Keilegom (2010) both estimate r by a local polynomial smoother $\hat{r}$ of degree d (see Section 2) and estimate $E(Y^2|X = x)$ by a local polynomial smoother $\hat{s}$ of degree d (defined as in Section 2 replacing Y_i with Y_i^2 for each $i = 1, \ldots, n$). Then, Neumeyer and Van Keilegom (2010) estimate $\sigma^2(x)$ by $\hat{\sigma}^2(x) = \hat{s}(x) - \{\hat{r}(x)\}^2$ at each x. The estimator $\hat{\sigma}$ for σ is then defined pointwise at each x as $\hat{\sigma}(x) = \sqrt{\hat{\sigma}^2(x)}$.

The estimators $\hat{r}$ and $\hat{\sigma}$ permit the desired expansion, if the assumptions of Theorem 3.1 (below) are satisfied. Theorem 2.1 of Neumeyer and Van Keilegom (2010) is proven under the following conditions:

ASSUMPTION 3.1. *The covariate vector X has a distribution that is quasi-uniform on the cube $[0,1]^m$; i.e. X has a density that is bounded and bounded away from zero on $[0,1]^m$. Additionally, all partial derivatives of the distribution function G of X up to order $2d+1$ exist on the interior of $[0,1]^m$.*

Further, we will require that both the regression function r and the scale function σ have partial derivatives of order $d+2$ (or higher), and that σ is non-vanishing. The choice of the degree d of the local polynomial smoothers will also depend on

certain smoothness and moment conditions and on the dimension of the covariate vector. In our simulation study we consider infinitely differentiable regression and scale functions and a one-dimensional covariate. This allows us to use locally linear smoothers.

THEOREM 3.1 (NEUMEYER AND VAN KEILEGOM (2010), THEOREM 2.1). *Let Assumption 3.1 be satisfied. Suppose the error distribution function F of ε is twice continuously differentiable, $\sup_{t\in\mathbb{R}} |tF''(t)| < \infty$ and $E(\varepsilon^6) < \infty$. Further, suppose that all partial derivatives of the functions r and σ up to order $d+2$ exist on the interior of $[0,1]^m$, they are uniformly continuous and σ is non-vanishing on $[0,1]^m$. Consider the estimators $\hat{r}$ and $\hat{\sigma}$ above with densities $w_1, \ldots, w_m$ supported on $[-1,1]$ that are symmetric, d-times continuously differentiable and, for $j = 1, \ldots, d-1$, $w_i^{(j)}(\pm 1) = 0$ for each $i = 1, \ldots, m$. Let the bandwidth c_n satisfy $nc_n^{2d+2} \to 0$ and $nc_n^{3m+\delta} \to \infty$ for some $\delta > 0$. Then, for $\hat{\varepsilon}_i = \{Y_i - \hat{r}(X_i)\}/\hat{\sigma}(X_i)$, $i = 1, \ldots, n$,*

$$\sup_{t\in\mathbb{R}} \Big| \frac{1}{n} \sum_{j=1}^{n} \Big[\mathbf{1}(\hat{\varepsilon}_j \le t) - \mathbf{1}(\varepsilon_j \le t) - f(t)\Big\{ \varepsilon_j + \frac{t}{2}(\varepsilon_j^2 - 1) \Big\} \Big] \Big| = o_p(n^{-1/2}).$$

The transfer principle for complete case statistics may now be used to adapt the results of Theorem 3.1 to the MAR model. To do this, we first identify the corresponding complete case estimator of F as

$$\hat{\mathbb{F}}_c = \frac{1}{N} \sum_{j=1}^{n} \delta_j \mathbf{1}(\hat{\varepsilon}_{j,c} \le t) = \frac{1}{N} \sum_{j=1}^{n} \delta_j \mathbf{1}\Big(\frac{Y_j - \hat{r}_c(X_j)}{\hat{\sigma}_c(X_j)} \le t \Big). \tag{3.2}$$

Here $\hat{r}_c$ and $\hat{\sigma}_c$ are the complete case versions of $\hat{r}$ and $\hat{\sigma}$, respectively, and each is defined analogously to (2.3) in Section 2. A requirement of the transfer principle is for the conditional joint distribution of (X, Y) given $\delta = 1$ to meet the assumptions of the joint distribution of (X, Y) imposed by Theorem 3.1. Similar to the observations of Section 2, it is easy to see this requirement only effects the covariate distribution

G. Hence, the complete case statistic $\hat{\mathbb{F}}_c$ is a consistent estimator for F in the MAR model. Thus only Assumption 3.1 must be restated.

ASSUMPTION 3.2. *The conditional distribution of the covariate vector X given $\delta = 1$ is quasi-uniform on the cube $[0,1]^m$; i.e. it has a density that is bounded and bounded away from zero on $[0,1]^m$. Additionally, all partial derivatives of the distribution function G_1 of X given $\delta = 1$ up to order $2d+1$ exist on the interior of $[0,1]^m$.*

By the transfer principle, the complete case version of the estimator from Theorem 3.1 satisfies expansion (3.1). This expansion is equivalent to

$$\sup_{t\in\mathbb{R}} \Big| \frac{1}{n} \sum_{j=1}^{n} \frac{\delta_j}{E\delta} \Big[\mathbf{1}(\hat{\varepsilon}_{j,c} \le t) - \mathbf{1}(\varepsilon_j \le t) - \Big\{ \varepsilon_j + \frac{t}{2}(\varepsilon_j^2 - 1) \Big\} \Big] \Big| = o_p(n^{-1/2}). \tag{3.3}$$

Hence, we have, uniformly in $t \in \mathbb{R}$,

$$\hat{\mathbb{F}}_c = \frac{1}{n} \sum_{j=1}^{n} \frac{\delta_j}{E\delta} \mathbf{1}(\hat{\varepsilon}_{j,c} \le t) + o_p(n^{-1/2}) = F(t) + \frac{1}{n} \sum_{j=1}^{n} b(\delta_j, \varepsilon_j, t) + o_p(n^{-1/2}),$$

where the function $b(\delta, \varepsilon, t) = \delta / E\delta [\mathbf{1}(\varepsilon \le t) - F(t) + f(t)\{\varepsilon + t/2(\varepsilon^2 - 1)\}]$ is the influence function for $\hat{\mathbb{F}}_c$.

For the second argument, we find the influence function of $\hat{\mathbb{F}}_c$ is the *efficient influence function* for estimating $F(t)$: see Corollary 3.1 below. We now state the main result of this section:

THEOREM 3.2. *Consider the heteroskedastic nonparametric regression model with responses missing at random. Suppose the assumptions of Theorem 3.1 are satisfied, with Assumption 3.2 in place of Assumption 3.1. Then the complete case estimator $\hat{\mathbb{F}}_c$ of the error distribution function F satisfies the uniform stochastic expansion (3.3); i.e.*

$$\sup_{t\in\mathbb{R}} \Big| \frac{1}{n} \sum_{j=1}^{n} \frac{\delta_j}{E\delta} \Big[\mathbf{1}(\hat{\varepsilon}_{j,c} \le t) - \mathbf{1}(\varepsilon_j \le t) - \Big\{ \varepsilon_j + \frac{t}{2}(\varepsilon_j^2 - 1) \Big\} \Big] \Big| = o_p(n^{-1/2}).$$

Furthermore, if the error density function satisfies Assumption 3.3, stated below, then $\hat{\mathbb{F}}_c(t)$ *is asymptotically efficient, for estimating* $F(t)$, $t \in \mathbb{R}$, *with influence function*

$$b(\delta, \varepsilon, t) = \frac{\delta}{E\delta}\Big[\mathbf{1}(\varepsilon \leq t) - F(t) + f(t)\Big\{\varepsilon + \frac{t}{2}(\varepsilon^2 - 1)\Big\}\Big].$$

As in Section 2, we note the uniform expansion above implies the existence of a functional central limit theorem. In addition to this, the property that $\hat{\mathbb{F}}_c$ is efficient means that competing estimators will not achieve smaller mean squared error for large samples. This includes estimators that employ imputation approaches to estimate the missing responses, which is analogous to the conclusions of Section 2. Therefore we recommend the use of the complete case estimator $\hat{\mathbb{F}}_c$ for conducting various hypothesis tests concerning the heteroskedastic model. Our simulation studies are similar to those of Section 2 and Müller et al. (2012), where a test for normal errors is conducted.

3.1 Efficiency

Here we will construct the efficient influence function for estimating a linear functional $E\{h(\varepsilon)\}$ based on observations of the form $(X, \delta Y, \delta)$. We will first follow the arguments of Section 2, which considers estimation of the error distribution function from a nonparametric regression with responses missing at random. In doing so we follow the arguments of Müller et al. (2006), who consider linear functionals of the joint distribution of X and Y with data of the above form. Then, we follow the arguments of Schick (1994), who consider estimation of functionals from various heteroskedastic regression models. We only summarize their main arguments and refer to their papers for more details. Finally, we focus on the functional $F(t)$, which is done in Section 2 for the nonparametric MAR model. This allows us to adapt part

of those proofs to the model considered here.

In the following, no assumption of a parametric model (finite dimensional) is imposed on any of the regression function, the scale function or the joint probability distribution of the observations. Thus, the parameter set Θ consists of the unknown functions of the statistical model: a family of covariate distributions, $\mathscr{G}$; a family of error distributions, $\mathscr{F}$; a set of regression functions, $\mathscr{R}$; a set of scale functions, $\mathscr{S}$; and a family of response probability distributions, $\mathscr{B}$. That is, $\Theta = \mathscr{G} \times \mathscr{F} \times \mathscr{R} \times \mathscr{S} \times \mathscr{B}$. We impose the following assumptions:

ASSUMPTION 3.3. *The error density f is absolutely continuous with almost everywhere continuous derivative f' and finite Fisher information for both location and scale; i.e.*

$$\int (1+z^2)\Big(\frac{f'(z)}{f(z)}\Big)^2 dF(z) < \infty.$$

Since the construction of the efficient influence function utilizes directional information in Θ, we now identify the set of perturbations $\dot{\Theta}$; which may be thought of as directions. Proceeding as in Section 2, the joint distribution $P(dx, dy, dz)$ takes the form

$$P(dx, dy, dz) = G(dx)B_{\pi(x)}(dz)\{zQ(x, dy) + (1-z)\delta_0(dy)\},$$

where $B_p = p\delta_1 + (1-p)\delta_0$ denotes the Bernoulli distribution with parameter p and δ_t as the Dirac measure at t. The model considered here deviates from that considered in Section 2 only in the conditional distribution of Y given $X = x$. Therefore, perturbations G_{nu} of G, π_{nw} of π and Q_{nv} of Q that are Hellinger differentiable require that u, w and v be restricted to subspaces of $\mathcal{L}_{2,0}(G)$, $\mathcal{L}_2(G_\pi)$ and $\mathcal{V}_0$, respectively.

In this work no model assumptions on G and π have been made. Thus, we only need to identify the appropriate subspace $\mathcal{V}$ of $\mathcal{V}_0$. Since the covariates and the errors

are assumed to be independent, we may write

$$Q(x, dy) = f\Big\{\frac{y - r(x)}{\sigma(x)}\Big\}\frac{1}{\sigma(x)}dy.$$

Using this notation, the constraint on $v(x, y) \in \mathcal{V}_0$ states that $\int v(x, y)f[\{y - r(x)\}/\sigma(x)]/\sigma(x)dy = 0$. In order to derive the explicit form of $\mathcal{V}$, we introduce further perturbations s, t and m of the unknown functions f, r and σ, respectively, and write

$$Q_{nv}(x, dy) = Q_{nstm}(x, dy) = f_{ns}\Big\{\frac{y - r_{nt}(x)}{\sigma_{nm}(x)}\Big\}\frac{1}{\sigma_{nm}(x)}dy,$$

where $f_{ns}(z) = f(z)\{1 + n^{-1/2}s(z)\}$, $r_{nt}(x) = r(x) + n^{-1/2}t(x)$ and $\sigma_{nm}(x) = \sigma(x) + n^{-1/2}m(x)$ for $s \in \mathcal{S}$, $t \in \mathcal{T}$ and $m \in \mathcal{M}$. Here

$$\mathcal{S} = \Big\{s \in \mathcal{L}_2(F) : \int s(z)f(z)dz = 0, \int zs(z)f(z)dz = 0 \text{ and } \int z^2s(z)f(z)dz = 0\Big\},$$

which is derived by the constraints that f_{ns} must integrate to one and have both a mean of zero and unit variance. This work assumes no parametric forms for r and σ. Hence, we take $\mathcal{T}$ to be $\mathcal{L}_2(G_1)$ and $\mathcal{M}$ to be $\mathcal{L}_2(G_1)$. In the following we will write "$\doteq$" to denote asymptotic equivalence; i.e. equality up to an additive term of order $o_p(n^{-1/2})$. Similar to the calculations of Section 2 and Schick (1994), who considers,

more generally, directionally differentiable regression and scale functions, we have

$$\begin{aligned}
f_{ns}\left\{\frac{y-r_{nt}(x)}{\sigma_{nm}(x)}\right\}\frac{1}{\sigma_{nm}(x)} &\doteq \left[f\left\{\frac{y-r_{nt}(x)}{\sigma_{nm}(x)}\right\}\right]\times\left[\frac{1}{\sigma(x)}\left\{1-n^{-1/2}\frac{m(x)}{\sigma(x)}\right\}\right] \\
&\quad\times\left[1+n^{-1/2}s\left\{\frac{y-r_{nt}(x)}{\sigma_{nm}(x)}\right\}\right] \\
&\doteq f\left\{\frac{y-r(x)}{\sigma(x)}\right\}\frac{1}{\sigma(x)}\times\left[1+n^{-1/2}\mathbf{k}^T(x)\ell\left\{\frac{y-r(x)}{\sigma(x)}\right\}\right] \\
&\quad\times\left[1+n^{-1/2}s\left\{\frac{y-r(x)}{\sigma(x)}\right\}\right] \\
&\doteq f\left\{\frac{y-r(x)}{\sigma(x)}\right\}\frac{1}{\sigma(x)} \\
&\quad\times\left(1+n^{-1/2}\left[\mathbf{k}^T(x)\ell\left\{\frac{y-r(x)}{\sigma(x)}\right\}+s\left\{\frac{y-r(x)}{\sigma(x)}\right\}\right]\right),
\end{aligned}$$

where $\ell(z)=(\ell_1(z),\ell_2(z))^T$, for $\ell_1(z)=-f'(z)/f(z)$ and $\ell_2(z)=-1-zf'(z)/f(z)$, and $\mathbf{k}(x)=(t(x)/\sigma(x),m(x)/\sigma(x))^T$. Note the equivalence of $t\in\mathcal{L}_2(G_1)$ and $m\in\mathcal{L}_2(G_1)$ with $\mathbf{k}\in\mathcal{L}_2(G_1)\times\mathcal{L}_2(G_1)$. Hence,

$$Q_{ns\mathbf{k}}(x,dy)\doteq f\left\{\frac{y-r(x)}{\sigma(x)}\right\}\frac{1}{\sigma(x)}\left(1+n^{-1/2}\left[\mathbf{k}^T(x)\ell\left\{\frac{y-r(x)}{\sigma(x)}\right\}+s\left\{\frac{y-r(x)}{\sigma(x)}\right\}\right]\right)$$

and $\mathcal{V}$ takes the form

$$\mathcal{V}=\left\{v(x,y)=\mathbf{k}^T(x)\ell\left\{\frac{y-r(x)}{\sigma(x)}\right\}+s\left\{\frac{y-r(x)}{\sigma(x)}\right\}:\mathbf{k}\in\mathcal{L}_2(G_1)\times\mathcal{L}_2(G_1),s\in\mathcal{S}\right\}.$$

Thus we specify the set $\dot{\Theta}$ as all Hellinger perturbations of the statistical model parameters; i.e. $\dot{\Theta}=\mathcal{L}_{2,0}(G)\times\mathcal{S}\times\{\mathcal{L}_2(G_1)\times\mathcal{L}_2(G_1)\}\times\mathcal{L}_2(G_\pi)$. For $\gamma=(u,s,\mathbf{k},w)$ in $\dot{\Theta}$, the perturbed distribution $P_{n\gamma}(dx,dy,dz)$ of an observation $(X,\delta Y,\delta)$ is then

$$P_{n\gamma}=G_{nu}(dx)B_{\pi_{nw}(x)}(dz)\{zQ_{nstm}(x,dy)+(1-z)\delta_0(dy)\}.$$

It follows that P is Hellinger differentiable with tangent

$$d_\gamma(X, \delta Y, \delta) = u(X) + \{\delta - \pi(X)\}w(X) + \delta\{\mathbf{k}^T(X)\ell(\varepsilon) + s(\varepsilon)\}.$$

Here we may take the tangent space T to be the closure of the linear subspace formed by d_γ; i.e.

$$T = \mathcal{L}_{2,0}(G) \oplus \big\{\{\delta - \pi(X)\}w(X) : w \in \mathcal{L}_2(G_\pi)\big\} \oplus \{\delta v(X, Y) : v \in \mathcal{V}\}.$$

We are interested in the linear functional $E\{h(\varepsilon)\}$. In order to specify a gradient for $E\{h(\varepsilon)\}$, we first need to find its directional derivative $\gamma_h \in \dot{\Theta}$, which is characterized by a limit as follows. As in Müller et al. (2004), we have, for every $s \in S$,

$$\lim_{n\to\infty} n^{1/2}\Big[\int h(z) f_{ns}(z)\, dz - \int h(z) f(z)\, dz\Big] = E\{h(\varepsilon)s(\varepsilon)\} = E\{h_0(\varepsilon)s(\varepsilon)\},$$

with h_0 given as a projection of h onto $\mathcal{S}$:

$$\begin{aligned} h_0(z) = h(z) &- E\{h(\varepsilon)\} - zE\{\varepsilon h(\varepsilon)\} \\ &- \frac{z^2 - E(\varepsilon^3)z - 1}{E(\varepsilon^4) - E^2(\varepsilon^3) - 1}\big[E\{\varepsilon^2 h(\varepsilon)\} - E(\varepsilon^3)E\{\varepsilon h(\varepsilon)\} - E\{h(\varepsilon)\}\big]. \end{aligned}$$

Thus, $E\{h(\varepsilon)\}$ is directionally differentiable with directional derivative $(0, h_0, \mathbf{0}, 0)$ and gradient $h_0(\varepsilon)$. By the convolution theorem (see, for example, Section 2 of Schick (1993)) the unique canonical gradient $g^*(X, \delta Y, \delta)$ is found by orthogonally projecting the gradient $h_0(\varepsilon)$ onto the tangent space T. Thus, $g^*(X, \delta Y, \delta)$ must take the form

$$g^*(X, \delta Y, \delta) = u^*(X) + \{\delta - \pi(X)\}w^*(X) + \delta\{\mathbf{k}^{*T}(X)\ell(\varepsilon) + s^*(\varepsilon)\}. \tag{3.4}$$

This yields the following projection equation

$$E\{h_0(\varepsilon)s(\varepsilon)\} = E\{g^*(X, \delta Y, \delta)d_\gamma(X, \delta Y, \delta)\}, \tag{3.5}$$

which must hold for all γ in $\dot{\Theta}$. A straightforward calculation of the right-hand side of (3.5) yields

$$\begin{aligned} E\{h_0(\varepsilon)s(\varepsilon)\} = E\{u^*(X)u(X)\} + E[\{\delta - \pi(X)\}^2 w^*(X)w(X)] \\ + E\{\delta \mathbf{k}^{*T}(X)\ell(\varepsilon)\ell^T(\varepsilon)\mathbf{k}(X)\} + E\{\delta \mathbf{k}^{*T}(X)\ell_0(\varepsilon)s^*(\varepsilon)\} \\ + E\delta E\{s^*(\varepsilon)s(\varepsilon)\}, \end{aligned}$$

where $\ell_0(\varepsilon)$ is a projection of ℓ onto $\mathcal{V}$; i.e. set

$$\ell_0(z) = \ell(z) - z\mathbf{e}_1 - \frac{z^2 - E(\varepsilon^3)z - 1}{E(\varepsilon^4) - E^2(\varepsilon^3) - 1}\{2\mathbf{e}_2 - E(\varepsilon^3)\mathbf{e}_1\},$$

with $\mathbf{e}_1 = (1,0)^T$ and $\mathbf{e}_2 = (0,1)^T$. For later calculational ease, we introduce the following quantities: $J = E\{\ell(\varepsilon)\ell^T(\varepsilon)\}$ and $J_0 = E\{\ell_0(\varepsilon)\ell_0^T(\varepsilon)\}$. Further, write

$$\ell_d(z) = \ell(z) - \ell_0(z) = z\mathbf{e}_1 + \frac{z^2 - E(\varepsilon^3)z - 1}{E(\varepsilon^4) - E^2(\varepsilon^3) - 1}\{2\mathbf{e}_2 - E(\varepsilon^3)\mathbf{e}_1\}$$

and J_d, which is calculated analogously to J, and J_0, and simplifies to $J_d = J - J_0$.

From (3.5) it is clear that $u^* = w^* \equiv 0$. Setting $u = w \equiv 0$ along with $\mathbf{k} \equiv \mathbf{0}$ in (3.5) we obtain

$$E\{h_0(\varepsilon)s(\varepsilon)\} = E\delta E[E_1\{\mathbf{k}^{*T}(X)\}\ell_0(\varepsilon)s(\varepsilon)] + E\delta E\{s^*(\varepsilon)s(\varepsilon)\}.$$

This implies

$$0 = E\left\{\left(s^*(\varepsilon) - \left[\frac{1}{E\delta}h_0(\varepsilon) - E_1\{\mathbf{k}^{*T}(X)\}\ell_0(\varepsilon)\right]\right)s(\varepsilon)\right\},$$

for every $s \in \mathcal{S}$. Thus, $s^*(z) = (E\delta)^{-1}h_0(z) - E_1\{\mathbf{k}^{*T}(X)\}\ell_0(z)$. Now set $u = w =$

$s \equiv 0$ in (3.5) and plug-in $s^*(\varepsilon)$, given above, to obtain

$$\begin{aligned}
0 &= E\delta E_1\{\mathbf{k}^{*T}(X)J\mathbf{k}(X)\} \\
&\quad + E\delta E\left(E_1\{\mathbf{k}^T(X)\}\ell_0(\varepsilon)\left[\frac{1}{E\delta}h_0(\varepsilon) - E_1\{\mathbf{k}^{*T}(X)\}\ell_0(\varepsilon)\right]\right) \\
&= E\delta E_1\{\mathbf{k}^{*T}(X)J\mathbf{k}(X)\} + E_1\{\mathbf{k}^T(X)\}E\{\ell_0(\varepsilon)h_0(\varepsilon)\} \\
&\quad - E\delta E_1\{\mathbf{k}^{*T}(X)\}J_0E_1\{\mathbf{k}(X)\}.
\end{aligned}$$

To continue, analogously to Müller et al. (2004) write $\mathcal{L}_2(G_1) \times \mathcal{L}_2(G_1)$ as an orthogonal sum of functions with mean zero and of constants; i.e. $\mathcal{L}_2(G_1) \times \mathcal{L}_2(G_1) = \{\mathcal{L}_{2,0}(G_1) \times \mathcal{L}_{2,0}(G_1)\} \oplus \{[1] \times [1]\}$, which means we may write $\mathbf{k}(x) = [\mathbf{k}(x) - E_1\{\mathbf{k}(X)\}] + E_1\{\mathbf{k}(X)\}$. Doing this, the above equation now becomes

$$\begin{aligned}
0 &= E\delta E_1\Big(\big[\mathbf{k}^*(X) - E_1\{\mathbf{k}^*(X)\}\big]^T J\big[\mathbf{k}(X) - E_1\{\mathbf{k}(X)\}\big]\Big) \\
&\quad + E_1\{\mathbf{k}^T(X)\}E\{\ell_0(\varepsilon)h_0(\varepsilon)\} - E\delta E_1\{\mathbf{k}^{*T}(X)\}J_dE_1\{\mathbf{k}(X)\},
\end{aligned}$$

for every $\mathbf{k} \in \mathcal{L}_2(G_1) \times \mathcal{L}_2(G_1)$. This implies

$$\mathbf{k}^* \equiv E_1\{\mathbf{k}^*(X)\},\ E_1\{\mathbf{k}^*(X)\} = -\frac{1}{E\delta}J_d^{-1}E\{\ell_0(\varepsilon)h_0(\varepsilon)\}$$

and, thus, $\mathbf{k}^* \equiv -(E\delta)^{-1}J_d^{-1}E\{\ell_0(\varepsilon)h_0(\varepsilon)\}$. Combining the above calculations we obtain to following result:

LEMMA 3.1. *The canonical gradient of $E\{h(\varepsilon)\}$ is $g^*(X, \delta Y, \delta)$ and is characterized by $(0, s^*, \mathbf{k}^*, 0)$, where*

$$s^*(z) = \frac{1}{E\delta}h_0(z) - E_1\{\mathbf{k}^{*T}(X)\}\ell_0(z) \text{ and } \mathbf{k}^* \equiv -\frac{1}{E\delta}J_d^{-1}E\{\ell_0(\varepsilon)h_0(\varepsilon)\},$$

with the quantities

$$h_0(z) = h(z) - E\{h(\varepsilon)\} - zE\{\varepsilon h(\varepsilon)\} \\ - \frac{z^2 - E(\varepsilon^3)z - 1}{E(\varepsilon^4) - E^2(\varepsilon^3) - 1}\big[E\{\varepsilon^2 h(\varepsilon)\} - E(\varepsilon^3)E\{\varepsilon h(\varepsilon)\} - E\{h(\varepsilon)\}\big],$$

$$\ell_0(z) = \ell(z) - z\mathbf{e}_1 - \frac{z^2 - E(\varepsilon^3)z - 1}{E(\varepsilon^4) - E^2(\varepsilon^3) - 1}\{2\mathbf{e}_2 - E(\varepsilon^3)\mathbf{e}_1\}$$

and

$$J_d^{-1} = \frac{1}{E(\varepsilon^4) - E^2(\varepsilon^3) - 1}\begin{bmatrix} E(\varepsilon^4) - 1 & -2E(\varepsilon^3) \\ -2E(\varepsilon^3) & 4 \end{bmatrix}$$

An estimator $\hat{\mu}$ for $E\{h(\varepsilon)\}$ is called efficient, in the sense of Hájek and Le Cam, if it is asymptotically linear with corresponding influence function equal to the canonical gradient $g^*(X, \delta Y, \delta)$ that characterizes $E\{h(\varepsilon)\}$; i.e. if

$$n^{1/2}\big\{\hat{\mu} - E\{h(\varepsilon)\}\big\} = n^{-1/2}\sum_{i=1}^{n} g^*(X_i, \delta_i Y_i, \delta_i) + o_p(1).$$

Combining this fact with Lemma 3.1 and (3.4) we obtain the following result:

THEOREM 3.3. *Consider the heteroskedastic nonparametric regression model with responses missing at random. An estimator $\hat{\mu}$ of $E\{h(\varepsilon)\}$ is efficient if it satisfies the expansion*

$$n^{1/2}\big\{\hat{\mu} - E\{h(\varepsilon)\}\big\} = n^{-1/2}\sum_{i=1}^{n}\frac{\delta}{E\delta}\Big(h_0(\varepsilon_i) - \big[E\{h_(\varepsilon_i)\ell_0(\varepsilon_i)\}\big]^T J_d^{-1}\ell_d(\varepsilon_i)\Big) + o_p(1),$$

where

$$\ell_d(z) = z\mathbf{e}_1 + \frac{z^2 - zE(\varepsilon^3) - 1}{E(\varepsilon^4) - E^2(\varepsilon^3) - 1}\{2\mathbf{e}_2 - E(\varepsilon^3)\mathbf{e}_1\}.$$

In this work we are interested in studying a residual-based goodness-of-fit test, which requires estimation of the error distribution function. Additionally, many statistical procedures are residual-based and rely on estimation of the error distribution

function. Thus, the function $h(z) = \mathbf{1}(z \leq t)$ is of particular interest. In the following result, using Theorem 3.3 with this $h(z)$, we obtain the expansion for an efficient residual-based estimator of the error distribution function:

COROLLARY 3.1. *Consider the heteroskedastic nonparametric regression model with responses missing at random. An estimator $\hat{F}$ of F is efficient, in the sense of Hájek and Le Cam, if it satisfies the expansion*

$$n^{1/2}\{\hat{F}(t) - F(t)\} = n^{-1/2} \sum_{i=1}^{n} \frac{\delta}{E\delta} \left[\mathbf{1}(\varepsilon_i \leq t) - F(t) + f(t) \left\{ \varepsilon_i + \frac{t}{2}(\varepsilon_i^2 - 1) \right\} \right] + o_p(1).$$

This corresponds to the expansion (3.3) of the complete case estimator $\hat{\mathbb{F}}_c$ from Section 3, which serves as the proof of Theorem 3.2.

3.2 Test for normal errors

To conclude this section we conduct a test for normal errors of a heteroskedastic nonparametric regression. In order to preserve the nonparametric nature of the unknown regression and scale functions, we choose for the simulations

$$r(x) = x^3 - x^2 + x + \cos\left(\frac{3\pi}{2}x\right) \quad \text{and} \quad \sigma(x) = \frac{1}{2} + \cos^2\left(\frac{\pi}{2}x\right).$$

The covariates were generated from a uniform distribution: $X_i \sim U(-1, 1)$ for $i = 1, \ldots, n$; see Figure 3.1 which shows a scatterplot of a simulated dataset. Finally, the indicators δ_i have a Bernoulli($\pi(x)$) distribution, with $\pi(x) = P(\delta = 1 | X = x)$. For the study, we use a logistic distribution function for $\pi(x)$, with a mean of 0 and a scale parameter of 1; see Section 2 for details. As a consequence, the average amount of missing data is around 50% ranging between 27% and 73%. We choose to work with $d = 1$, the locally linear smoother, using bandwidth $c_n = 1.25\{n \log(n)\}^{-1/4}$. The assumptions of Theorem 3.2 are then satisfied for the choices made above.

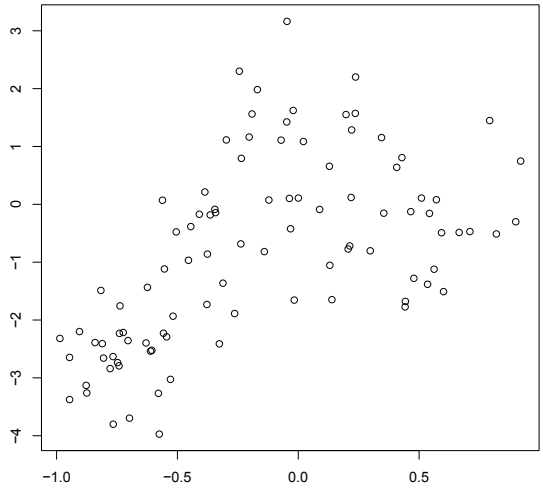

Figure 3.1: $r(x) = x^3 - x^2 + x + \cos\left(\frac{3\pi}{2}x\right)$, $\sigma(x) = \frac{1}{2} + \cos^2\left(\frac{\pi}{2}x\right)$ $-1 \leq x \leq 1$, with $N(0,1)$ errors

Several approaches are known in the literature to test for normal errors. That is, using the density function f, we test the null hypothesis

$$H_0 : f(x) = \phi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}, \qquad x \in \mathbb{R}.$$

For example, one could use the Kolmogorov-Smirnov test statistic

$$T_{KS} = \sup_{t \in \mathbb{R}} \left| \hat{\mathbb{F}}_c(t) - \Phi(t) \right|,$$

which seeks to find the largest deviation between the empirical distribution function and the standard normal distribution function. A popular approach is to consider the Cramér-von Mises statistic

$$T_{CvM} = n \int_0^\infty \left\{ \hat{\mathbb{F}}_c(u) - \Phi(u) \right\}^2 \phi(u) du,$$

which is based on a distance between the empirical distribution function and the standard normal distribution function.

We prefer using a martingale transform approach. The technique was studied in the MAR model in Section 4 of Koul et al. (2012), who consider a semiparametric model. Our approach is related to a special case of that above. Here the linear term is zero and the regression function is estimated using a locally linear smoother. Writing

$$h(x) = (1, x, x^2 - 1)^T = (1, -\phi'(x)/\phi(x), -(x\phi'(x))/\phi(x))^T,$$

$$\Gamma(x) = \int_x^\infty h(u)h^T(u)\phi(u)du,$$

and

$$H(t) = \int_{-\infty}^t h^T(s)\Gamma^{-1}(s)\phi(s)ds,$$

the test statistic of Koul et al. (2012) is given by

$$T_{MT1} = \sup_{t\in\mathbb{R}} \Big| \frac{1}{\sqrt{N}} \sum_{j=1}^n \delta_j \big\{ \mathbf{1}(\hat{\varepsilon}_{j,c} \le t) - H(t \wedge \hat{\varepsilon}_{j,c}) h(\hat{\varepsilon}_{j,c}) \big\} \Big|.$$

Using the transfer principle, Koul et al. (2012) argue the complete case version of the test statistic under the MAR model has the same limiting distribution as that of the full model. In order to conduct the test using their test statistic, Koul et al. (2012) state that estimators of the error distribution function should satisfy

$$\sup_{t\in\mathbb{R}} \Big| \frac{1}{\sqrt{N}} \sum_{j=1}^n \delta_j \Big[\mathbf{1}(\hat{\varepsilon}_{j,c} \le t) - \mathbf{1}(\varepsilon_j \le t) - f(t) \Big\{ \varepsilon_j + \frac{t}{2}(\varepsilon_j^2 - 1) \Big\} \Big] \Big| = o_p(1),$$

when the scale of the errors is unknown. This expansion is equivalent to (3.3).

A possible alternative to the approach using the standardized residuals $\hat{\varepsilon}_c$ is to use the standardized residuals $\hat{Z}_c$, which are constructed using a different scale function estimator. Here, estimate $\mathrm{Var}(Y|X = x, \delta = 1) = E_1[\{Y - r(x)\}^2|X = x]$ by a local polynomial smoother $\hat{\tau}_c$ of degree of d. It is defined analogously to (2.3) in

Test for normal errors								
	$N(0,1)$		$(\chi^2_1-1)/\sqrt{2}$		$t_5/\sqrt{5/3}$		Laplace$(0,1)$	
n	T_{MT1}	T_{MT2}	T_{MT1}	T_{MT2}	T_{MT1}	T_{MT2}	T_{MT1}	T_{MT2}
50	0.069	0.045	0.072	0.081	0.069	0.054	0.062	0.046
200	0.082	0.064	0.997	0.995	0.110	0.099	0.175	0.157

Table 3.1: Simulated level ($N(0,1)$ figures) and power for T_{MT1}, T_{MT2}

Section 2 with Y_i replaced by $\{Y_i - \hat{r}_c(X_i)\}^2$ for each $i = 1, \ldots, n$. In particular, we can alternatively estimate the scale function $\sigma(x)$ by $\sqrt{\hat{\tau}_c(x)}$ at each x. The standardized residuals $\hat{Z}_c$ are then defined similarly to $\hat{\varepsilon}_c$; i.e.

$$\hat{Z}_{i,c} = \frac{Y_i - \hat{r}_c(X_i)}{\sqrt{\hat{\tau}_c(X_i)}}, \qquad \text{for } i = 1, \ldots, n.$$

Then we define the test statistic T_{MT2} as T_{MT1} above, but plugging in $Z_{i,c}$ for $\hat{\varepsilon}_{i,c}$ for each $i = 1, \ldots, n$.

A simulation was conducted using 1000 runs at sample sizes 50 and 200. The critical value for the test is given by the upper 5% quantile of the $\sup_{t\in[0,1]} |B(t)|$ distribution, where $B(t)$ is the standard Brownian motion, which is approximately 2.2414. First we consider the (5%) level of the test by inspecting the results for the $N(0,1)$ distribution. In order to check the power of the test, we consider several alternative error distributions: a mean shifted and rescaled $\chi^2(1)$ distribution, a rescaled $t(5)$ distribution, and a Laplace distribution with mean 0 and variance 1.

Table 3.1 shows that when the errors are normally distributed (the null hypothesis is true) the test using T_{MT1} rejects the null hypothesis 6.9% of the time for samples of size 50, and 8.2% of the time for samples of size 200. We suspect this liberal behavior is attributable to a slight bias that occurs when constructing T_{MT1}. Since the estimator $\hat{\sigma}^2_c$ is negative for some values of x, we calculated its absolute value before taking the square root in constructing the estimator $\hat{\sigma}_c$. The results using T_{MT2}

are similar to those of T_{MT1}, but slightly less in value. When the null hypothesis is not true, the power results are similar between both tests. Here, the test using T_{MT1} appears to be slightly more powerful at small sample sizes than the test using T_{MT2}. The differences between the two tests are less pronounced at the larger sample size of 200. In conclusion, both test procedures seem to have similar performance.

4. A DISTRIBUTION FREE TEST FOR HETEROSKEDASTIC ERRORS IN NONPARAMETRIC REGRESSION WITH MISSING DATA

An important assumption made in regression is that variation in the data remains constant across values of the covariates. That is, we wish to study the model

$$Y = r(X) + \sigma_0 \xi,$$

which we call the *homoskedastic model*. Here the function r is called the regression function, X and $\varepsilon = \sigma_0 \xi$ are independent and σ_0 is a strictly positive-valued constant. When the assumption of constant variation is relaxed the variation in the data no longer remains constant across values of the covariates. This is called *heteroskedasticity*, and thus we study the model

$$Y = r(X) + \sigma(X) \xi,$$

which we call the *heteroskedastic model*. Here the function σ is called the scale function, which varies with the covariates X. Studying the difference between homoskedastic and heteroskedastic models is important because it can be difficult to determine which model is appropriate in practice. If the heteroskedasticity is not properly handled, then many statistical procedures will lead to inconsistent results. Thus, testing for its presence is of great importance in many statistical analyses.

The two models above are related. They represent the conclusions of the following

statistical hypotheses:

$$H_0 : \exists \sigma_0 > 0,\ \sigma(x) = \sigma_0 \quad a.e.(G)$$
$$H_a : \sigma(\cdot) \in \Sigma,$$

where $\Sigma = \{\sigma \in \mathcal{L}_2(G) : \sigma(\cdot) > 0 \text{ and non-constant } a.e.(G)\}$ is a space of scale functions. Here G is the distribution function of the covariates X. When the null hypothesis is true, the homoskedastic model is appropriate for the data. However, when the alternative hypothesis is true, the heteroskedastic model is appropriate for the data. Rejection of the null hypothesis would imply that sufficient statistical evidence is gathered in the data to declare the homoskedastic model inappropriate.

Remark 4.1. Observe that $\sigma_0^2 = E\{\sigma^2(X)\}$ and that testing the above hypotheses are logically equivalent to testing the hypotheses, for $h(x) = \sigma(x)/\sigma_0$,

$$H_0 : h(x) = 1 \quad a.e.(G)$$
$$H_a : h(\cdot) \in \Sigma',$$

where

$$\Sigma' = \{h \in \mathcal{L}_2(G) : h(x) > 0 \text{ and non-constant } a.e.(G),\ E\{h(X)\} = 1\}$$

is a space of functions. The function h may be thought of as the essence of the heteroskedasticity in the model. Thus, as above, rejection of the null hypothesis would imply that there is sufficient statistical evidence in the data to declare the homoskedastic model inappropriate. As a consequence, conclusions made concerning heteroskedasticity do not depend on the average amount of variation in the errors; i.e. σ_0 does not need to be estimated.

To estimate the regression function r we will use a nonparametric model. We

are concerned with the case where the responses Y are missing. This means that we observe a sample of data $(X_1, \delta_1 Y_1, \delta_1), \ldots, (X_n, \delta_n Y_n, \delta_n)$, where δ is an indicator taking values one, when Y is observed, and zero, otherwise. In this work we assume the responses are missing at random (MAR); see Section 2 for a definition.

In this section, we show the test statistic T_c, defined in equation (4.6) below, may be used to test for the presence of heteroskedasticity in the nonparametric MAR model. Our test statistic T_c uses only the completely observed data; i.e. we use only observations of the form (X, Y) called the complete cases. In particular, we use only the available residuals $\hat{\varepsilon}_{i,c} = Y_i - \hat{r}_c(X_i)$, where $\hat{r}_c$ is a suitable complete case estimator of the regression function r. Demonstrating this will require two steps.

First, we study the case where all indicators are equal to one, the *full model*. We will require the following condition:

ASSUMPTION 4.1. *The covariate vector X is quasi-uniform on the cube $[0,1]^m$; i.e. X has a density that is bounded and bounded away from zero on $[0,1]^m$.*

Now we will introduce some notation. We estimate the regression function r by a local polynomial smoother $\hat{r}$ of degree d; see Section 2 for a definition.

The estimator $\hat{r}$ permits the desired properties of Lemma 1 of Müller et al. (2012) (written as Lemma 4.1 below) when its assumptions are satisfied. This will require the regression function to be in the Hölder space $H(d, \gamma)$; i.e. it has continuous partial derivatives of order up to d (or higher) and the partial derivatives of order d are Hölder with exponent γ. Additionally, the error density f must satisfy certain smoothness and moment conditions. The choice of the degree d of the local polynomial smoother will depend on the dimension of the covariate vector.

LEMMA 4.1 (MÜLLER ET AL. (2009), LEMMA 1). *Let the distribution G of the covariates X satisfy Assumption 4.1. Suppose the regression function r belongs to*

the Hölder space $H(d,\gamma)$ with $s = d + \gamma > 3p/2$, the error variable has mean zero and a finite moment of order $\zeta > 4s/(2s-p)$, and the densities $w_1, \ldots, w_p$ are $(p+2)-$times continuously differentiable and have compact support $[-1,1]$. Let $c_n \sim \{n\log(n)\}^{-1/(2s)}$. Then there is a random function $\hat{a}$ such that

$$P(\hat{a} \in H_1(p,\alpha)) \to 1 \tag{4.1}$$

for some $\alpha > 0$,

$$\int |\hat{a}(x)|^{1+b} G(dx) = o_p(n^{-1/2}) \tag{4.2}$$

for $b > p/(2s-p)$,

$$\int \hat{a}(x) G(dx) = \frac{1}{n}\sum_{j=1}^{n} \varepsilon_j + o_p(n^{-1/2}), \tag{4.3}$$

and

$$\sup_{x\in\mathbb{R}^p} |\hat{r}(x) - r(x) - \hat{a}(x)| = o_p(n^{-1/2}). \tag{4.4}$$

The tests proposed in this section are inspired by those of Koul et al. (2012), who develop tests for linearity of a semiparametric regression in both the full model and the MAR model. These approaches are in the spirit of Stute (1997), who studies a weighted (marked) empirical process related to the integrated regression function. The resulting test statistic is based on a weighted empirical distribution function, which is strikingly simple. This gives the motivation for our test statistics below.

For the full model, the test statistic is given by

$$T_n = \sup_{t\in\mathbb{R}} \Big| \frac{1}{n}\sum_{j=1}^{n} W_j \mathbf{1}(\hat{\varepsilon}_j \le t) \Big| = \sup_{t\in\mathbb{R}} \Big| \frac{1}{n}\sum_{j=1}^{n} W_j \mathbf{1}(Y_j - \hat{r}(X_j) \le t) \Big|. \tag{4.5}$$

Here, the weights W are constructed by first choosing some $\omega \in \Sigma$. Then $W_i =$

$\{\omega(X_i) - n^{-1}\sum_{k=1}^n \omega(X_k)\}/[n^{-1}\sum_{j=1}^n \{\omega(X_j) - n^{-1}\sum_{k=1}^n \omega(X_k)\}^2]^{-1/2}$, for each $i = 1, \ldots, n$. We now state the main result for the test statistic T_n of the fully observed model.

THEOREM 4.1. *Let the null hypothesis hold. Suppose the assumptions of Lemma 4.1 are satisfied for the local polynomial smoother $\hat{r}$. Choose $\omega \in \Sigma$ and write $W_i = \{\omega(X_i) - n^{-1}\sum_{k=1}^n \omega(X_k)\}/[n^{-1}\sum_{j=1}^n \{\omega(X_j) - n^{-1}\sum_{k=1}^n \omega(X_k)\}^2]^{1/2}$ and $\hat{\varepsilon}_i = Y_i - \hat{r}(X_i)$ for $i = 1, \ldots, n$. Then*

$$T_n = \sup_{t\in\mathbb{R}} \Big| \frac{1}{\sqrt{n}} \sum_{j=1}^n W_j \mathbf{1}(\hat{\varepsilon}_j \le t) \Big|$$

converges in distribution to $\sup_{t\in[0,1]} |B_0(t)|$, where B_0 denotes the standard Brownian bridge.

For our second step, we now apply the transfer principle for complete case statistics (given in Koul et al. (2012)) to adapt the results of Theorem 4.1 to the MAR model. The complete case test statistic is given by

$$T_c = \sup_{t\in\mathbb{R}} \Big| \frac{1}{N} \sum_{j=1}^n \delta_j W_{j,c} \mathbf{1}(\hat{\varepsilon}_{j,c} \le t) \Big| = \sup_{t\in\mathbb{R}} \Big| \frac{1}{N} \sum_{j=1}^n \delta_j W_{j,c} \mathbf{1}(Y_j - \hat{r}_c(X_j) \le t) \Big|. \quad (4.6)$$

Here $N = \sum_{j=1}^n \delta_j$ is the number of complete cases and, for each $i = 1, \ldots, n$, $W_{i,c} = \{\delta_i\omega(X_i) - N^{-1}\sum_{k=1}^n \delta_k\omega(X_k)\}/[N^{-1}\sum_{j=1}^n \{\delta_j\omega(X_j) - N^{-1}\sum_{k=1}^n \delta_k\omega(X_k)\}^2]^{1/2}$. The estimator $\hat{r}_c$ is the corresponding complete case estimator to $\hat{r}$; see (2.3) of Section 2.

In order to use the transfer principle, the conditional distribution of (X, Y) given $\delta = 1$ must meet the assumptions of the joint distribution of (X, Y) imposed by Theorem 4.1. It is easy to see this will only affect the covariate distribution G. This may be seen by combining the MAR assumption with the independence of X and ε and observing that ε and (X, δ) are independent. Hence, the error distribution

function F and the functions ω and r remain the same when moving from the unconditional distribution to the conditional distribution. Thus, only Assumption 4.1 must be restated.

ASSUMPTION 4.2. *The conditional distribution G_1 of the covariate vector X given $\delta = 1$ is quasi-uniform on the cube $[0,1]^m$; i.e. it has a density that is bounded and bounded away from zero on $[0,1]^m$.*

The transfer principle implies the limiting behavior of the complete case test statistic of the MAR model is that of the corresponding test statistic of the full model. This means that T_c and T_n have the same limiting distribution. Combining these arguments provides proof for the main result of this section:

THEOREM 4.2. *Let the null hypothesis hold. Suppose the assumptions of Theorem 4.1 are satisfied, with Assumption 4.2 in place of Assumption 4.1. Write $W_{i,c} = \{\delta_i \omega(X_i) - N^{-1}\sum_{k=1}^n \delta_k \omega(X_k)\}/[N^{-1}\sum_{j=1}^n \{\delta_j \omega(X_j) - N^{-1}\sum_{k=1}^n \delta_k \omega(X_k)\}^2]^{1/2}$ and $\hat{\varepsilon}_{i,c} = Y_i - \hat{r}_c(X_i)$ for $i = 1, \ldots, n$. Then*

$$T_c = \sup_{t \in \mathbb{R}} \left| \frac{1}{\sqrt{N}} \sum_{j=1}^n \delta_j W_{j,c} \mathbf{1}(\hat{\varepsilon}_{j,c} \leq t) \right|$$

converges in distribution to $\sup_{t\in[0,1]} |B_0(t)|$, where B_0 denotes the standard Brownian bridge.

We note that the limiting distribution of a standard Brownian bridge provides both T_n and T_c with the property of being asymptotically distribution free. This means that inference made on heteroskedasticity using either T_n or T_c does not depend on unknown model parameters, specifically the error distribution function. Below, we conduct a small simulation study to demonstrate the effectiveness of conducting a hypothesis test using T_c for the nonparametric MAR model.

4.1 Auxiliary results

Here we derive the auxiliary results surrounding the test statistic T_n for the full model. We are interested in a weighted empirical process of the errors, and its corresponding approximation. Weighted empirical processes have been well studied in the literature. In particular, many useful results are stated in Koul (2002), which we will use in the following arguments and refer the reader there for further details. We now restate the setup given in page 28 of Koul (2002) in the following assumption:

ASSUMPTION 4.3. *Let $(\Omega, \mathscr{A}, P)$ be a probability space, and F be a distribution function on $\mathbb{R}$. For each integer $n \geq 1$, let $(\varepsilon_{ni}, W_{ni}, a_{ni})$, for $1 \leq i \leq n$, be an array of independent trivariate random variables defined on $(\Omega, \mathscr{A})$ such that $\varepsilon_{n1}, \ldots, \varepsilon_{nn}$ are independent and identically distributed according to F and independent of (W_{ni}, a_{ni}) for each i. Let $\mathscr{A}_{ni}$ be a filtration for each $1 \leq i \leq n$; i.e. $A_{n1} \subset A_{n2} \subset \cdots \subset A_{ni}$. Here (W_{n1}, a_{n1}) are $\mathscr{A}_{n1}-$measurable. Additionally, for each $1 \leq i \leq n$, the random variables $\varepsilon_{n1}, \ldots, \varepsilon_{n,i-1}$ and $(W_{n1}, a_{n1}), \ldots, (W_{ni}, a_{ni})$ are $\mathscr{A}_{ni}-$measurable with ε_{ni}*

independent of $\mathscr{A}_{ni}$. *Define the processes below as*

$$\hat{U}_n(t) = \frac{1}{n}\sum_{j=1}^{n} W_{nj}\mathbf{1}(\varepsilon_{nj} \leq t + a_{nj}),$$

$$\hat{J}_n(t) = \frac{1}{n}\sum_{j=1}^{n} E\{W_{nj}\mathbf{1}(\varepsilon_{nj} \leq t + a_{nj})|\mathscr{A}_{nj}\},$$

$$U_n(t) = \frac{1}{n}\sum_{j=1}^{n} W_{nj}\mathbf{1}(\varepsilon_{nj} \leq t),$$

$$J_n(t) = \frac{1}{n}\sum_{j=1}^{n} E\{W_{nj}\mathbf{1}(\varepsilon_{nj} \leq t)|\mathscr{A}_{nj}\},$$

$$\hat{V}_n(t) = n^{-1/2}\sum_{j=1}^{n} W_{nj}[\mathbf{1}(\varepsilon_{nj} \leq t + a_{nj}) - E\{W_{nj}\mathbf{1}(\varepsilon_{nj} \leq t + a_{nj})|\mathscr{A}_{nj}\}],$$

$$V_n(t) = n^{-1/2}\sum_{j=1}^{n} W_{nj}\{\mathbf{1}(\varepsilon_{nj} \leq t) - F(t)\}.$$

The process $V_n(t)$ is called a weighted empirical process and has an approximation $\hat{V}_n(t)$. Under very general conditions, Theorem 2.2.4 of Koul (2002) (written as Theorem 4.3 below) states that a weighted empirical process and its corresponding approximation are indistinguishable for large samples, and, hence, attain the same limiting distribution. In addition to this, the limiting distribution is known to be a product between some random variable and the standard Brownian bridge. This will require certain limiting probability conditions, for each i, on deviations a_{ni} and the weights W_{ni}, and the distribution function F must satisfy certain smoothness conditions.

Theorem 4.3 (Theorem 2.2.4 of Koul (2002)). *Let Assumption 4.3 hold. As-*

sume the following conditions hold:

$$\max_i |a_{ni}| = o_p(1),$$
$$n^{-1/2} \sum_{j=1}^{n} |W_{nj} a_{nj}| = O_p(1),$$

F has uniformly continuous a.e. positive Lebesgue density f,

$$\left| \frac{1}{n} \sum_{j=1}^{n} W_{nj}^2 \right|^{1/2} = \alpha + o_p(1), \quad \textit{for some positive r.v. } \alpha.$$

Then

$$\sup_{t \in \mathbb{R}} |\hat{V}_n(t) - V_n(t)| = o_p(1).$$

Further, assume the following conditions hold:

$$\mathscr{A}_{ni} \subset \mathscr{A}_{n+1,i} \quad 1 \leq i \leq n \quad n \geq 1,$$
$$W_{ni} \textit{ are square integrable} \quad 1 \leq i \leq n \quad n \geq 1.$$

Then, in distribution, as $n \to \infty$,

$$\hat{V}_n(t) \to \alpha B_0(F) \quad \textit{and} \quad V_n(t) \to \alpha B_0(F),$$

where B_0 is the standard Brownian bridge on $\mathscr{C}[0,1]$, independent of α.

We are interested in the difference between the two weighted empirical processes of Assumption 4.3, and, in particular, when faced with the deviation $\hat{a}$ between the true regression and the estimated regression. To study the difference between these processes, we will require the distribution function F of the errors to satisfy certain smoothness conditions that are more strict than those of Theorem 4.3. In addition, the deviations $\hat{a}$ are correlated. This will require that they tend in probability to lie on a function space of deviates a that depend only on the covariates X. Using these conditions, we obtain the following result:

LEMMA 4.2. *Consider the homoskedastic nonparametric regression model. Let the conditions of Assumption 4.3 hold with the additional condition that F has a density f that is (uniformly) Hölder continuous with exponent β. Further, let d be a non-negative function in $\mathcal{L}_2(G)$, and let $\mathscr{D}$ be a family of functions a so that $|a| < d$ and $0 \in \mathscr{D}$. Assume there is a function $\hat{a}$ that satisfies*

$$P(\hat{a} \in \mathscr{D}) \to 1 \quad \text{as } n \to \infty. \tag{4.7}$$

Write, for each $i = 1, \ldots, n$, $a_{ni} = \hat{a}(X_i)\mathbf{1}(\hat{a} \in \mathscr{D})$ and assume that

$$\max_i |a_{ni}| = o_p(1). \tag{4.8}$$

Further, assume that $W_{n1}, \ldots, W_{nn}$ are i.i.d. random variables that satisfy the following conditions:

$$\frac{1}{n}\sum_{j=1}^{n} |W_{nj}a_{nj}| = O_p(n^{-1/2}), \tag{4.9}$$

$$\frac{1}{n}\sum_{j=1}^{n} |W_{nj}||a_{nj}|^{1+\beta} = o_p(n^{-1/2}), \tag{4.10}$$

and, for some positive-valued random variable S,

$$\Big|\frac{1}{n}\sum_{j=1}^{n} W_{nj}^2\Big|^{1/2} = S + o_p(1). \tag{4.11}$$

Then,

$$\sup_{t\in\mathbb{R}} \Big|\frac{1}{n}\sum_{j=1}^{n} W_{nj}\big[\mathbf{1}\{\varepsilon_{nj} \le t + \hat{a}(X_j)\} - \mathbf{1}(\varepsilon_{nj} \le t) - f(t)\hat{a}(X_j)\big]\Big| = o_p(n^{-1/2}). \tag{4.12}$$

Proof. Our proof follows the style of the proof of Theorem 2.2 of Müller et al. (2007), and we refer the reader to that paper for further details. Since the event $\{\hat{a} \notin \mathscr{D}\}$ is

tending to zero in probability, it is sufficient to show

$$\sup_{t\in\mathbb{R}}\Big|\frac{1}{n}\sum_{j=1}^{n}W_{nj}\{\mathbf{1}(\varepsilon_{nj}\leq t+a_{nj})-\mathbf{1}(\varepsilon_{nj}\leq t)-f(t)a_{nj}\}\Big|=o_p(n^{-1/2}).$$

To do this, first write

$$W_n(t)=\frac{1}{n}\sum_{j=1}^{n}W_{nj}\{\mathbf{1}(\varepsilon_{nj}\leq t)-F(t)\}$$

and

$$\hat{W}_n(t)=\frac{1}{n}\sum_{j=1}^{n}W_{nj}\{\mathbf{1}(\varepsilon_{nj}\leq t+a_{nj})-F(t+a_{nj})\},$$

and consider the difference $\hat{W}_n(t)-W_n(t)$. The assumptions of Theorem 4.3 hold such that $\sup_{t\in\mathbb{R}}|\hat{W}_n(t)-W_n(t)|$ becomes

$$\sup_{t\in\mathbb{R}}\Big|\frac{1}{n}\sum_{j=1}^{n}W_{nj}\{\mathbf{1}(\varepsilon_{nj}\leq t+a_{nj})-\mathbf{1}(\varepsilon_{nj}\leq t)-F(t+a_{nj})+F(t)\}\Big|$$

and

$$\sup_{t\in\mathbb{R}}|\hat{W}_n(t)-W_n(t)|=o_p(n^{-1/2}). \tag{4.13}$$

Now consider the difference

$$\sup_{t\in\mathbb{R}}\Big|\frac{1}{n}\sum_{j=1}^{n}W_{nj}\{\mathbf{1}(\varepsilon_{nj}\leq t+a_{nj})-\mathbf{1}(\varepsilon_{nj}\leq t)-f(t)a_{nj}\}-\hat{W}_n(t)+W_n(t)\Big|$$

and find that it is equal to

$$\sup_{t\in\mathbb{R}}\Big|\frac{1}{n}\sum_{j=1}^{n}W_{nj}\{F(t+a_{nj})-F(t)-f(t)a_{nj}\}\Big|,$$

which is bounded by

$$\sup_{t\in\mathbb{R}}\frac{1}{n}\sum_{j=1}^{n}|W_{nj}||F(t+a_{nj})-F(t)-f(t)a_{nj}|.$$

Since f is Hölder with exponent β and (4.10), there is a constant C such that

$$\frac{1}{n}\sum_{j=1}^{n}|W_{nj}||F(t+a_{nj})-F(t)-f(t)a_{nj}| \leq C\frac{1}{n}\sum_{j=1}^{n}|W_{nj}||a_{nj}|^{1+\beta} = o_p(n^{-1/2}).$$

Note, this does not depend on t. Therefore,

$$\sup_{t\in\mathbb{R}}\left|\frac{1}{n}\sum_{j=1}^{n}W_{nj}\{F(t+a_{nj})-F(t)-f(t)a_{nj}\}\right| = o_p(n^{-1/2}). \tag{4.14}$$

The desired result (4.12) now follows by combining (4.13) with (4.14). □

To construct the test statistic, we must place certain conditions on the weights. In particular, we will require the last term in the left-hand side of (4.12) to vanish in the limit, which requires (4.10) and the density function f to be smooth. In addition to this, we will construct our weights to sum to zero and to have a standard deviation of one. We can then establish that our weighted empirical distribution function approximates a weighted empirical process that has nontrivial limiting behavior. This conclusion is given by the distributional results of Theorem 4.3. By adding these assumptions we obtain the following result:

PROPOSITION 4.1. *Let the assumptions of Lemma 4.2 hold, but replace condition (4.11) with*

$$\sum_{j=1}^{n}W_{nj}=0 \quad \textit{and} \quad \left|\frac{1}{n}\sum_{j=1}^{n}W_{nj}^2\right|^{1/2} = 1+o_p(1), \quad i=1,\ldots,n. \tag{4.15}$$

Then

$$\sup_{t\in\mathbb{R}}\left|\frac{1}{\sqrt{n}}\sum_{j=1}^{n}W_{nj}\mathbf{1}\{\varepsilon_{nj}\leq t+\hat{a}(X_j)\}\right| \to \sup_{t\in[0,1]}|B_0(t)|, \tag{4.16}$$

in distribution, as $n\to\infty$, where B_0 denotes the standard Brownian bridge.

Proof. Since the event $\{\hat{a}\notin\mathscr{D}\}$ is tending to zero in probability, it is sufficient to show the result for a_{ni}, $i=1,\ldots,n$. This is done in two steps. For the first step, the

assumptions of Lemma 4.2 are satisfied such that

$$\sup_{t\in\mathbb{R}}\Big|\frac{1}{n}\sum_{j=1}^{n}W_{nj}\big\{\mathbf{1}(\varepsilon_{nj}\le t+a_{nj})-\mathbf{1}(\varepsilon_{nj}\le t)-f(t)a_{nj}\big\}\Big|=o_p(n^{-1/2}). \tag{4.17}$$

Now consider the difference of the left-hand side of (4.17) from

$$\sup_{t\in\mathbb{R}}\Big|\frac{1}{n}\sum_{j=1}^{n}W_{nj}\big\{\mathbf{1}(\varepsilon_{nj}\le t+a_{nj})-\mathbf{1}(\varepsilon_{nj}\le t)\big\}\Big|,$$

which becomes

$$\sup_{t\in\mathbb{R}}\Big|\frac{1}{n}\sum_{j=1}^{n}W_{nj}f(t)a_{nj}\Big|$$

and is bounded by

$$\sup_{t\in\mathbb{R}}\frac{1}{n}\sum_{j=1}^{n}|W_{nj}f(t)a_{nj}|.$$

Since f is Hölder with exponent β and (4.10), there is a constant C such that

$$\frac{1}{n}\sum_{j=1}^{n}|W_{nj}f(t)a_{nj}|\le C\frac{1}{n}\sum_{j=1}^{n}|W_{nj}||a_{nj}|^{1+\beta}+o_p(n^{-1/2})=o_p(n^{-1/2}).$$

Note, this does not depend on t. Thus,

$$\sup_{t\in\mathbb{R}}\Big|\frac{1}{n}\sum_{j=1}^{n}W_{nj}f(t)a_{nj}\Big|=o_p(n^{-1/2})$$

and

$$\sup_{t\in\mathbb{R}}\Big|\frac{1}{n}\sum_{j=1}^{n}W_{nj}\big\{\mathbf{1}(\varepsilon_{nj}\le t+a_{nj})-\mathbf{1}(\varepsilon_{nj}\le t)\big\}\Big|=o_p(n^{-1/2}).$$

Now use (4.15) to calculate

$$\sup_{t\in\mathbb{R}}\Big|\frac{1}{n}\sum_{j=1}^{n}W_{nj}\big\{\mathbf{1}(\varepsilon_{nj}\le t+a_{nj})-\mathbf{1}(\varepsilon_{nj}\le t)+F(t)\big\}\Big|=o_p(n^{-1/2}). \tag{4.18}$$

For the second step, let i be given and consider an event A_i from the field $\mathscr{A}_{ni}$. In our case it is easy to see that A_i is also an event in $\mathscr{A}_{n+1,i}$ because, for the nonparametric regression model, information is added when moving from a sample

of size n to a sample of size $n+1$. This means that at least as much knowledge is gained by sampling $n+1$ individuals as there is when sampling n individuals. Thus, the conditions of Theorem 4.3 are satisfied such that

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n} W_{nj}\big\{\mathbf{1}(\varepsilon_{nj} \le t) - F(t)\big\}, \qquad \text{for } -\infty \le t \le \infty, \tag{4.19}$$

converges in $D([-\infty,\infty])$ to a time-changed Brownian bridge $B_0(F)$. The desired result (4.16) follows by combining (4.18) with (4.19). □

We may now construct our test statistic using a local polynomial smoother to estimate the unknown regression function. Let $\omega \in \Sigma$. Hence, for $i = 1,\ldots,n$, writing

$$W_i = \frac{\omega(X_i) - n^{-1}\sum_{j=1}^{n}\omega(X_j)}{\left[\frac{1}{n}\sum_{j=1}^{n}\{\omega(X_j) - n^{-1}\sum_{k=1}^{n}\omega(X_k)\}^2\right]^{1/2}},$$

(4.15) is satisfied and

$$\frac{1}{n}\sum_{j=1}^{n}|W_j| = O_p(n^{-1/2}).$$

Under the conditions of Lemma 4.1 the local polynomial estimator $\hat{r}$ admits a random function $\hat{a}$ so that (4.2) holds. Taking $\mathscr{D} = H_1(p,\alpha)$ together with (4.1) and (4.2) we find

$$\int |\hat{a}(x)|^{1+b}\mathbf{1}(\hat{a} \in \mathscr{D})G(dx) = o_p(n^{-1/2}), \tag{4.20}$$

for $b > p/(2s-p)$. Now use this fact in combination with Markov's inequality to find that (4.8) holds. To establish (4.9), first use (4.20) and observe that

$$E\{W_1\hat{a}(X_1)\mathbf{1}(\hat{a} \in \mathscr{D})\}^2 = \int \hat{a}^2(x)\mathbf{1}(\hat{a} \in \mathscr{D})G(dx) = o_p(n^{-1/2}),$$

and then use a law of large numbers. Equation (4.10) can be shown by using the

Markov inequality and (4.20). From the conclusion of Proposition 4.1 we have that

$$T_n = \sup_{t\in\mathbb{R}} \Big| \frac{1}{\sqrt{n}} \sum_{j=1}^{n} W_j \mathbf{1}(\hat{\varepsilon}_j \le t) \Big|, \qquad \text{for } -\infty \le t \le \infty,$$

converges in $D([-\infty, \infty])$ to a time-changed Brownian bridge $B_0(F)$. The above arguments serve as the proof of Theorem 4.1.

4.2 Simulation studies

To conclude this section we conduct a brief computational study of the previous results. To preserve the nonparametric nature of the unknown regression function, we choose for the simulations

$$r(x) = 2x^2 + 3\cos(\pi x).$$

The covariates were generated from a uniform distribution and errors from a standard normal distribution: $X_i \sim U(-1, 1)$ and $\varepsilon_i \sim N(0, 1)$ for $i = 1, \ldots, n$. Finally, the indicators δ_i have a Bernoulli($\pi(x)$) distribution, with $\pi(x) = P(\delta = 1|X = x)$. For the study, we use a logistic distribution function for $\pi(x)$ with a mean of 0 and a scale parameter of 1; see Section 2 for details. As a consequence, the average amount of missing data is around 50% ranging between 27% and 73%. We work with $d = 1$, the locally linear smoother, with bandwidth $c_n = 2\{n \log(n)\}^{-1/4}$, and $\omega(x) = \phi(x)$, with ϕ as the density function of the standard normal distribution. The assumptions of Theorem 4.2 are then satisfied for the choices made above. In order to investigate the level and power of the test, we consider the following scale functions for $x \in \mathbb{R}$:

$$\sigma_1(x) = 1, \qquad \sigma_2(x) = 1 + 2x^2,$$

$$\sigma_3(x) = 1 + \cos^2(\pi x), \qquad \sigma_4(x) = 1 + 4\frac{|x|}{\sqrt{n}}.$$

Test for heteroskedastic errors								
	T_n				T_c			
n	50	100	500	1000	50	100	500	1000
σ_1	0.007	0.013	0.040	0.061	0.000	0.001	0.003	0.006
σ_2	0.056	0.313	0.998	1.000	0.008	0.043	0.908	1.000
σ_3	0.000	0.006	0.055	0.193	0.001	0.002	0.015	0.049
σ_4	0.019	0.023	0.064	0.118	0.000	0.003	0.012	0.009

Table 4.1: Simulated level (σ_1 figures) and power for T_n and T_c

The critical value for the 5% level test is given by the upper 5% quantile of the $\sup_{t\in[0,1]} |B_0(t)|$ distribution, which is approximately 1.224. Here, the scale function σ_1 allows for the (5%) level of each test to be checked. The scale functions σ_2, σ_3 and σ_4 each give an indication of the power of each test in different scenarios.

Our simulation of 1000 runs was conducted using sample sizes 50, 100, 500, and 1000. Table 4.1 shows that when the scale function σ_1 is used (the null hypothesis is true) the test using T_n rejects the null hypothesis 0.7% of the time for samples of size 50 and 6.1% of the time for samples of size 1000. For the test using T_c, the null hypothesis was never incorrectly rejected for samples of size 50 but 0.6% of the time for samples of size 1000.

Inspecting the power of the test, for the σ_2 figures, the test using T_n appears to perform well rejecting the null hypothesis 31.3% of the time for samples of size 100 and all of the time for samples of size 1000. Similar findings occur for the test using T_c. For the σ_3 figures, both test procedures have poor performance. When no data are missing, rejection of the null hypothesis occurs 19.3% of the time at samples of size 1000. The results are worse when data are missing. Inspecting the σ_4 figures gives similar findings to those of σ_3. In conclusion, we find that the each test performs well in certain situations.

5. CONCLUSIONS

In conclusion, there is a vast literature guiding the construction of new nonparametric techniques, and the applicability of current nonparametric approaches in Statistics. In some cases, these methods offer alternatives to parametric methods, and, in cases where no parametric models are appropriate, these techniques may still be used. The ideas used in the previous sections rely heavily on specialized knowledge of those problems. For example, using the estimation efficiency criterion in Sections 2 and 3, and the weighted empirical process theory in Section 4. It is interesting for the popular assumptions of normally distributed errors and of homoskedastic errors each have nonparametric tests with similar properties as their parametric counterparts. Naturally, this leads to the question of finding which assumptions, particularly those guided by applications of Statistics, can be tested by such nonparametric approaches.

From Section 2, we find the question of normal errors is answerable by a nonparametric technique. Our estimator is shown to be optimal in the sense that it is least disperse among the class of regular estimators (those estimators that have limiting distributions). As a result, this is counter-intuitive to conventional wisdom because the approach ignores the data that are not fully observed. When the technique of González-Manteiga and Pérez-González (2006), which we refer to as double smoothing, was applied only minor gains were observed in power for smaller sample sizes. For example, the complete case test statistic T_c lead to incorrect rejection of the null hypothesis 2.2% of the time and to correct rejection of mean shifted $\chi^2(1)$ errors 48.9% of the time at a sample size of 50, and the imputed test statistic T_ι lead to incorrect rejection of the null hypothesis 2.5% of the time and correct rejection of

the null hypothesis of mean shifted $\chi^2(1)$ errors 53.5% of the time at a sample size of 50. This is a more dramatic example of the importance of choosing estimators based on optimality criterion.

The results in Section 3 highlight the importance of focusing on key aspects of questions of interest to Statistics. Viewing the question of normally distributed errors as a goodness-of-fit problem embedded in a heteroskedastic model revealed a technique that paralleled that of Section 2. This is in stark contrast to the duality of wisdoms between fully observed and missing data environments. As a consequence, the results discovered in Section 3 mirror very closely those achieved in Section 2, but this is a positive result for applications of Statistics. Interestingly, it means that, in some cases, intuition may be taken from homoskedastic models and applied to heteroskedastic models, even with missing data.

Finally, in Section 4, we find the question of the existence of heteroskedasticity is answerable using a nonparametric approach. Our technique produces a test statistic that is asymptotically distribution free. However, simulations revealed the test was performing poorly for some heteroskedastic models. This leads to the question of how much improvement could be gained by changing the weights used in those simulations. In all, this problem highlights the fact that even nonparametric approaches can be influenced by an observer of science, and, therefore, could be subject to certain experimenter biases.

REFERENCES

D. Asteriou and S. G. Hall. *Applied Econometrics*. Palgrave MacMillan, New York, New York, 2011. ISBN 9780230271821.

J. Chown and U.U. Müller. Efficiently estimating the error distribution in nonparametric regression with responses missing at random. *Journal of Nonparametric Statistics*, 25:665–677, 2013.

S. Efromovich. Nonparametric regression with responses missing at random. *Journal of Statistical Planning and Inference*, 141:3744–3752, 2011a.

S. Efromovich. Nonparametric regression with predictors missing at random. *Journal of the American Statistical Association*, 106:306–319, 2011b.

W. González-Manteiga and A. Pérez-González. Goodness-of-fit tests for linear regression models with missing response data. *Canadian Journal of Statistics*, 34: 149–170, 2006.

W.H. Greene. *Econometric Analysis*. Prentice Hall, Upper Saddle River, New Jersey, 2000. ISBN 9780130132970.

J.D. Hart. *Nonparametric Smoothing and Lack-of-fit Tests*. Springer Series in Statistics. Springer, New York, New York, 1997. ISBN 9780387949802.

E.V. Khmaladze and H.L. Koul. Martingale transforms goodness-of-fit tests in regression models. *Annals of Statistics*, 32:995–1034, 2004.

E.V. Khmaladze and H.L. Koul. Goodness-of-fit problem for errors in nonparametric regression: distribution free approach. *Annals of Statistics*, 37:3165–3485, 2009.

H.L. Koul. *Weighted Empirical Processes in Dynamic Nonlinear Models*. Lecture Notes in Statistics. Springer, New York, New York, 2002. ISBN 9780387954769.

H.L. Koul, U.U. Müller, and A. Schick. The transfer principle: a tool for complete case analysis. *Annals of Statistics*, 40:3031–3049, 2012.

H. Liang, S. Wang, and R. Carroll. Partially linear models with missing response variables and error-prone covariates. *Biometrika*, 94:185–198, 2007.

R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, Hoboken, New Jersey, 2002. ISBN 9780471183860.

G. Molenberghs and M. Kenward. *Missing Data in Clinical Studies*. Statistics in Practice. Wiley, Hoboken, New Jersey, 2007. ISBN 9780470849811.

U.U. Müller. Estimating linear functionals in nonlinear regression with responses missing at random. *Annals of Statistics*, 37:2245–2277, 2009.

U.U. Müller, A. Schick, and W. Wefelmeyer. Estimating linear functionals of the error distribution in nonparametric regression. *Journal of Statistical Planning and Inference*, 119:75–93, 2004.

U.U. Müller, A. Schick, and W. Wefelmeyer. Imputing responses that are not missing. In M. Nikulin, D. Commenges, and C. Huber, editors, *Probability, Statistics and Modelling in Public Health*, pages 350–363. Springer, New York, New York, 2006. ISBN 9780387260235.

U.U. Müller, A. Schick, and W. Wefelmeyer. Estimating the error distribution in semiparametric regression. *Statistics and Decisions*, 25:1–18, 2007.

U.U. Müller, A. Schick, and W. Wefelmeyer. Estimating the error distribution function in nonparametric regression with multivariate covariates. *Statistics and Probability Letters*, 79:957–964, 2009.

U.U. Müller, A. Schick, and W. Wefelmeyer. Estimating the error distribution function in semiparametric additive regression models. *Journal of Statistical Planning and Inference*, 142:552–566, 2012.

N. Neumeyer and I. Van Keilegom. Estimating the error distribution in nonparametric multiple regression with applications to model testing. *Journal of Multivariate Analysis*, 101:1067–1078, 2010.

D. Ruppert, M.P. Wand, and R.J. Carroll. Semiparametric regression during 2003–2007. *Electronic Journal of Statistics*, 3:1193, 2009.

A. Schick. On efficient estimation in regression models. *Annals of Statistics*, 21: 1486–1521, 1993.

A. Schick. On efficient estimation in regression models with unknown scale functions. *Mathematical Methods of Statistics*, 3:171–212, 1994.

S. J. Sheather. *A Modern Approach to Regression with R*. Springer Series in Language and Communication. Springer, New York, New York, 2009. ISBN 9780387096087.

W. Stute. Nonparametric model checks for regression. *Annals of Statistics*, 25: 613–641, 1997.

A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer, New York, New York, 2006. ISBN 9780387324487.

H.D. Vinod. *Hands-on Intermediate Econometrics Using R: Templates for Extending Dozens of Practical Examples*. World Scientific, Hackensack, New Jersey, 2008. ISBN 9789812818850.